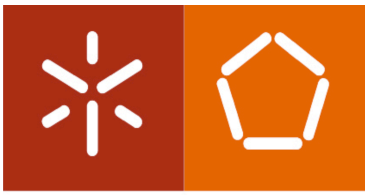

Universidade do Minho
Escola de Engenharia

Carlos Daniel Dias Correia

Data Mining e *Data Quality* em Dados da Saúde

Dissertação de Mestrado

Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação de

Professor Doutor Manuel Filipe Santos

Professor Doutor Carlos Filipe Portela

Outubro de 2017

DECLARAÇÃO

Nome: Carlos Daniel Dias Correia

Endereço eletrónico: carlosddcorreia@gmail.com

Telefone: 917490982

Número do Bilhete de Identidade: 14294245

Título da dissertação: *Data Mining* e *Data Quality* em Dados da Saúde

Orientadores:

- Professor Doutor Manuel Filipe Santos
- Professor Doutor Carlos Filipe Portela

Ano de conclusão: 2017

Designação do Mestrado: Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação

1. É AUTORIZADA A REPRODUÇÃO PARCIAL DESTA DISSERTAÇÃO, APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE.

Universidade do Minho, 31 de Outubro de 2017

Assinatura:

_____

(Carlos Daniel Dias Correia)

AGRADECIMENTOS

Ao Professor Doutor Manuel Filipe Santos, orientador de mestrado, pela orientação na realização deste projeto de investigação.

Ao Professor Doutor Carlos Filipe Portela, coorientador de mestrado, um especial agradecimento por toda a entrega e disponibilidade, mas acima de tudo por todo interesse e todo o conhecimento partilhado ao longo da realização do projeto.

Aos meus pais pelo esforço, pelo apoio incondicional, mas fundamentalmente por todo o investimento que fizeram na concretização de um sonho, a formação académica. Por serem modelos de coragem e ajuda na superação de dificuldades ao longo desta caminhada.

Às minhas "estrelas" avó e irmã, que marcaram e muito o meu percurso académico e pessoal.

À minha família por todo o apoio incondicional e por todos os ensinamentos.

À Paula pelo carinho, pelo conforto, pela ajuda, pelo conhecimento transmitido, pelos conselhos, pela dedicação, pela confiança, pelo mimo e acima de tudo por ser a melhor madrinha.

Ao Rui pelo ídolo que sempre foi, por influenciar diretamente e ativamente no meu crescimento, por todos os momentos inesquecíveis proporcionados, por todas as brincadeiras e fundamentalmente por ser como um irmão.

À Juliana pela paciência, apoio e carinho. Pela confiança transmitida, pela valorização tão entusiasta do meu trabalho dando-me, desta forma coragem para ultrapassar todos os obstáculos. Por ter sido o meu refúgio em todos os momentos.

À minha terra e a todas as pessoas maravilhosas que sempre me conheceram, acompanhando todo o meu percurso académico. Por me terem acolhido e apoiado em todas as fases difíceis da minha vida. Obrigado pela amizade, por me ensinarem a dar valor, pelo companheirismo e sobretudo pela felicidade que me proporcionaram.

Ao meu grupo de amigos, que sempre me deram força e coragem. Um especial obrigado por todas as aventuras e gargalhadas. Por todas as discussões, por todas as visões partilhadas, por todos os conselhos, por todas as conversas, por todos os desabafos e por todo conforto.

A todos os colegas de equipa, por todo o companheirismo, confiança e motivação. São sem dúvida alguma, marcos no meu crescimento partilhando momentos marcantes e inesquecíveis. Todas as derrotas, todas as vitórias, todos os berros, todos os conselhos, todas as dicas, todas as palmadas nas costas e por todas as palavras sábias em momentos mais difíceis.

A todos, um profundo e sincero obrigado!

Resumo

Este trabalho enquadra-se no desenvolvimento do projeto de dissertação de mestrado em Engenharia e Gestão de Sistemas de Informação da Universidade do Minho, sobre o tema – *Data Mining* e *Data Quality* em dados da Saúde. Este tema surge da interação entre um grupo de Investigação da Universidade do Minho e a Entidade Reguladora da Saúde.

Atualmente qualquer assunto relacionado com a saúde é sempre um tema muito sensível na sociedade, já que interfere diretamente no bem-estar das pessoas. Neste sentido, com o intuito de melhorar a qualidade dos serviços de saúde, é fundamental uma boa gestão da qualidade das reclamações.

Devido ao volume de reclamações, surge a necessidade de exploração de modelos de *Data Science* com o intuito de automatizar processos internos de qualidade das reclamações. Assim, o objetivo principal deste projeto passa pela melhoria da qualidade do processo de análise das reclamações na saúde, bem como a análise de conhecimento ao nível dos sistemas de informação aplicados à referida saúde.

Este documento desenvolvido no âmbito da dissertação de mestrado tem como objetivo apresentar uma contextualização do tema, bem como a motivação do desenvolvimento do mesmo.

Numa primeira fase do desenvolvimento deste projeto foi adquirido conhecimento científico útil no contexto do problema e também nas áreas de *Data Mining* e *Data Science*. De forma a garantir uma melhor resposta a este caso de estudo sentiu-se a necessidade de estudar e conjugar três metodologias, nomeadamente *Case Study*, *Design Science Research Methodology* e *Kimball Lifecycle*. No mesmo contexto, foi ainda realizado um estudo de viabilidade da aplicação de modelos de *Data Mining* através da metodologia Crisp-DM.

Numa segunda fase do desenvolvimento deste projeto foram adquiridos conhecimentos relativos aos dados fornecidos para estudo assim como desenhado e pensado todo o processo de desenvolvimento da solução. É observável o desenvolvimento do tratamento dos dados em duas etapas: carregamento dos dados para uma base de dados auxiliar e tratamento dos mesmos através do processo de *Extract, Transform e Load* (ETL). Com o *data warehouse* criado foi desenvolvido o cubo *Online Analytical Processing* (OLAP) que posteriormente foi interligado no *Power BI* possibilitando a criação e análise de *dashboards.*

Palavras-Chave

Data Mining, Data Science, Descoberta de Conhecimento em Base de Dados, Sistemas de Informação na Saúde, *Business Intelligence*, Qualidade das Reclamações na Saúde.

Abstract

This project is part of the development of the master thesis in Engineering and Management of Information Systems from the University of Minho, about the subject – Data Mining and Data Quality in health data. This subject occurs from the interaction between a research group from the University of Minho and the Health Regulatory Entity.

Nowadays, any health-related issue is always a very sensitive issue in the society as it interferes directly in the people well-being. In this sense, in order to improve the quality of health services, a good quality management of complaints is essential.

Due to the volume of complaints, there is a need to explore Data Science models in order to automate internal quality complaints processes. Thus, the main objective of this project is to improve the quality of the health claims analysis process, as well as the knowledge analysis at the level of information systems applied to referred health.

This document developed within the scope of the master thesis has as the main objective the presentation of a contextualization of the theme, as well as the motivation of the development of the theme in question.

In the first phase of the development of this project was acquired useful scientific knowledge in the context of the problem and also in the areas of Data Mining and Data Science. In order to guarantee a better response to this case study, it was necessary to study and combine three methodologies, namely Case Study, Design Science Research Methodology and Kimball Lifecycle. In the same context, a feasibility study of the application of Data Mining models through the Crisp-DM methodology was also performed.

In a second phase of the development of this project were acquired knowledge about the data provided for study as well as designed and thought the entire process of developing the solution. It is observable the development of data treatment in two stages: loading the data to an auxiliary database and processing them through the Extract, Transform and Load (ETL) process. With the data warehouse created, the Online Analytical Processing (OLAP) cube was developed that was later interconnected in Power BI enabling the creation and analysis of dashboards.

Keywords

Data Mining, Data Science, Knowledge Discovery in Database, Health Information Systems, Business Intelligence, Quality of Health Complaints.

ÍNDICE

Índice de Figuras

Índice de Tabelas

Lista de Abreviaturas, Siglas e Acrónimos

BI – Business Intelligence

CRISP-DM – Cross Industry Standard Process for Data Mining

DCBD – Descoberta de Conhecimento em Base de Dados

DM – Data Mining

DS – Data Science

DSRM – Design Science Research Methodology

DSS – Decision Support Systems

DW – Data Warehouse

ERS – Entidade Reguladora da Saúde

FNAM – Federação Nacional dos Médicos

NPM – New Public Management

OF- Ordem dos Farmacêuticos

OM – Ordem dos Médicos

PEM – Prescrição Eletrónica Médica

PPP – Parcerias Público Privado

SAD – Sistema de Apoio à Decisão

SAM – Sistema de Apoio ao Médico

SAPE – Sistema de Apoio à Prática de Enfermagem

SGQ – Sistema de Gestão da Qualidade

SI – Sistema de Informação

SNS – Serviço Nacional de Saúde

SPMS – Serviços Partilhados do Ministério da Saúde

ETL – Extract, Transform, Load

SSIS – SQL Server Integration Services

SSAS – SQL Server Analysis Services

OLAP – Online Analytical Processing

1 INTRODUÇÃO

Este capítulo encontra-se dividido em quatro seções sendo elas o enquadramento e motivação, entidade reguladora da saúde, objetivos e resultados e estruturação do documento. Na secção referente ao enquadramento e motivação é apresentada uma abordagem geral da importância deste projeto de investigação. Na secção seguinte é feita uma apresentação da entidade reguladora da saúde evidenciando a sua importância neste contexto. Na subsequente secção são descritos os objetivos que devem ser atingidos bem como os resultados expectáveis. Por último, na secção de estruturação do documento é apresentada a modelação do documento bem como a identificação do conteúdo de cada capítulo.

1.1 Enquadramento e Motivação

Hoje em dia é fundamental uma boa gestão da qualidade da informação por parte das organizações. Através de uma sistematização consistente dos dados é possível criar padrões que posteriormente poderão ser moldáveis dando respostas às necessidades do cliente, permitindo maior acesso e perceção da informação gerada. Com o avanço das tecnologias é possível conceber padrões com recurso a técnicas de *Data Mining (DM)* e *Data Science* (DS).

Existem vários estudos que procuram satisfazer as necessidades dos centros hospitalares no que toca à organização e qualidade de toda a informação gerada. Uma das soluções passa pela adoção de um sistema de gestão da qualidade (SGQ) que favoreça a qualidade da informação e agregue valor aos resultados em saúde (Freixo & Rocha, 2014).

No que toca ao serviço de saúde é importante mencionar o Serviço Nacional de Saúde (SNS) e a Entidade Reguladora da Saúde (ERS) como grandes entidades, que de alguma forma vão regularizando e supervisionando os estabelecimentos prestadores de cuidados de saúde.

No decorrer desta dissertação é expectável aprofundar conhecimentos científicos para que seja possivel identificar os problemas que existem relativamente aos processos internos de qualidade nos centros hospitalares.

Os sistemas de informação têm vindo a exercer um papel fundamental no que toca à saúde, pois permitem uma melhor organização e qualidade da informação. Todas as informações registadas em centros hospitalares podem ser essenciais em problemas com futuros utentes. Com o avanço das tecnologias, é possível melhorar os processos internos de elevada importância dos centros hospitalares.

Hoje em dia, a qualidade dos dados é uma das maiores preocupações das entidades da Saúde. Nesse sentido, pretende-se explorar técnicas de *Data Mining* e *Data Science* de modo a automatizar os processos internos de qualidade e auditoria das reclamações.

1.2 Entidade Reguladora da Saúde

Desde a década de 90, Portugal tem sofrido algumas mudanças no que toca à base principal da vida, a Saúde. As iniciativas reformistas foram variadas, sempre no sentido de melhorar essencialmente a gestão hospitalar (Simões, 2004). Segundo Reis (2011), os serviços prestadores de saúde têm vindo a estar sujeitos a mudanças ao nível da gestão, em busca de eficiência e qualidade dos serviços prestados. No seguimento destas mudanças surge um novo conceito de gestão - Novo Sistema de Gestão Pública (NPM)[1], sendo adotado pelo governo para melhorar a gestão hospitalar. Com mudanças neste setor, o modelo tradicional de gestão hospitalar adotado começou a ser incapaz de satisfazer as ideias estabelecidas para o serviço de saúde. Estas iniciativas reformistas da saúde, influenciadas pelo NPM, surgem da aplicação de princípios empresariais com missão social, às organizações públicas prestadoras de serviços de saúde, tornando-se um poderoso instrumento de mudança (A. S. Ferreira, 2004). Assim, em 2002 surge a empresarialização de hospitais e Parcerias Público Privado (PPP)[2]. Com a empresarialização hospitalar, o governo pretendia atribuir mais responsabilidade focando-se na conversão de centros hospitalares em empresas públicas de modo a obter resultados mais satisfatórios no que toca à eficiência e rentabilidade (V. Moreira, 2011; Reis, 2011).

No sentido de melhorar cada vez mais a regulação e monitorização dos estabelecimentos prestadores de saúde, surge em 2003, a criação da Entidade Reguladora de Saúde. Esta apresenta-se como uma entidade pública e independente, dotada para servir os utentes essencialmente a níveis de segurança, qualidade e direitos (Anabela, 2014). A ERS deparou-se com graves problemas pouco depois da sua criação, isto porque esta nova entidade era malvista por muita gente. Segundo V. Moreira (2011), " a ERS venceu entretanto o teste do tempo e o teste da legitimidade, só restando a oposição intransigente, atávica e sectária da ordem dos médicos".

Após as várias queixas e o desagrado de muitas pessoas, a ERS conseguiu prevalecer e com auxílio de algumas alterações nas legislações conseguiu ainda organizar-se em quatro departamentos fundamentais (L. Almeida, 2010):

- Departamento de Gestão Interna, responsável pela gestão administrativa e gestão de recursos humanos;

[1] (NPM) – *New Public Management*, utilizado por empresas e instituições para enfatizar o conceito de que as regras de gestão utilizadas no setor privado têm uma eficácia superior às usadas no setor público. (Reis, 2011)

[2] (PPP) - Parcerias Público Privado, baseia-se num contrato de gestão envolvendo atividades desde a conceção, construção, financiamento, conservação e exploração dos ativos infraestruturais até à gestão geral do hospital (Simões, 2004).

- Departamento de Proteção da Qualidade e Direitos dos Cidadãos, responsável por assegurar os direitos dos utentes bem como os processos de qualificação dos estabelecimentos prestadores de saúde;
- Departamento de Acompanhamento do Sistema de Saúde e Defesa do Acesso e da Concorrência, responsável pela restrição de acessos ao sistema e o supervisionamento da concorrência do mercado administrativo da saúde;
- Departamento de Supervisão e intervenção Jurídica, responsável pelo cumprimento da legislação e das respetivas sanções.

Segundo Almeida (2010), a implementação da ERS suscitou o desagrado de alguns parceiros sociais como a Ordem dos Médicos (OM), Ordem dos Farmacêuticos (OF) e da Federação Nacional dos Médicos (FNAM), com opiniões claramente críticas focando-se essencialmente na independência da ERS. Estas opiniões foram suavizadas, nomeadamente pela OM, contudo a falta de apoios institucionais prevalecia contribuindo para a demissão, em 2005, do presidente da ERS Rui Nunes.

A ERS nos seus primeiros anos de atuação no setor da Saúde atravessou momentos muito difíceis e mesmo sem apoios por parte das instituições já existentes, conseguiu prevalecer mantendo a sua missão. Esta missão identifica-se pela regulação da atividade dos estabelecimentos prestadores de cuidados de saúde, assegurando a sua supervisão e funcionamento no que diz respeito ao cumprimento dos requisitos estabelecidos por lei. Assim, pretendia-se assegurar a legalidade, a transparência e os direitos dos utentes face às relações económicas entre os diversos operadores (L. Almeida, 2010; Entidade Reguladora da Saúde, 2016).

A ERS está preparada para responder e ajudar nas mais variadas questões reguladoras, contudo não estando satisfeitos, tem apresentado uma evolução notória ao longo dos anos. Segundo L. Almeida (2010), a ERS tem desenvolvido muitas iniciativas inovadoras, nomeadamente:

- Diagnóstico da qualidade dos serviços públicos de saúde;
- Avaliação dos cuidados de saúde primários;
- Análise das queixas e reclamações dos utentes;
- Carta dos direitos do utente dos serviços de saúde;
- Sistema de registo das entidades reguladoras;
- Sistema de avaliação em saúde;
- Deteção de práticas de indução artificial da procura;
- Deteção das práticas de seleção de doentes;
- Avaliação de práticas de transferência e referenciação dos doentes;

- Regime de licenciamento dos estabelecimentos prestadores dos cuidados de saúde;
- Regime das convenções celebrados pelo SNS;
- Caracterização dos centros de nascimento não públicos;
- Análise da concorrência no sector do transporte de doentes;
- Análise da concorrência no sector da hemodiálise da informação através da criação e gestão da página *online* da ERS.

A ERS procura acentuar a sua eficácia na regulação da saúde em duas vertentes fundamentais, a económica e a social, sendo que a vertente económica lida essencialmente com o controlo de preços, produção e mercado, já a vertente social lida essencialmente com o controlo do cumprimento dos direitos dos utentes (L. Almeida, 2010).

1.3 Objetivos e resultados

A saúde é uma área muito complexa, o que leva à necessidade de se adquirir conhecimentos referentes a alguns conceitos fundamentais para uma boa interação com o tema proposto. Assim, foram abordados conceitos como a ERS e o SNS, percebendo o seu papel no que toca à qualidade e auditoria da informação. Um tema que também foi abordado é os Sistemas de Informação (SI) e qual o seu papel nos centros hospitalares. Juntamente a estas pesquisas, foram identificadas aplicações automatizadas, percebendo de que forma estas funcionalidades melhoram a qualidade da saúde.
Deste modo, este projeto de dissertação prende-se à seguinte questão científica:

- *De que modo os modelos de Data Mining podem melhorar a avaliação da qualidade e auditoria em saúde?*

Em virtude da especificidade do tema, da pertinência do trabalho a realizar e dos dados disponibilizados a questão de investigação teve de ser reestruturada para:

- *De que modo Data Science pode melhorar a qualidade do processo de análise das reclamações na saúde?*

De seguida, como principal objetivo, foram explorados modelos de DS para exploração e melhoria das diferentes dimensões da qualidade dos dados automatizando processos. Pretende-se uma melhoria dos padrões de resposta bem como uma análise de conhecimento ao nível dos Sistemas de Informação (SI) aplicados à saúde. No mesmo contexto, foi ainda realizado um estudo de viabilidade da aplicação de modelos de DM.

Em consonância com a questão científica e o objetivo principal advêm alguns objetivos secundários na realização deste projeto de dissertação:

- Estudo do negócio em causa;
- Análise e tratamento dos dados recebidos;
- Criação de modelos, com base nas informações recolhidas;
- Desenvolvimento de *dashboards* para visualização e análise da informação;
- Criação de modelos de classificação e regressão.

No desenvolvimento deste projeto de dissertação foram ainda consideradas e aplicadas técnicas como *Business Intelligence* e *Big Data*. Desta forma, o BI foi aplicado com o intuito de análise e tratamento dos dados e demonstração dos resultados, como por exemplo em dashboards. Por outro lado, o *Big Data* foi aplicado no que toca à quantidade de dados registados e o processamento dos mesmos em tempo real.

1.4 Estrutura do documento

Este documento está dividido em 5 grande capítulos. O capítulo 1 consiste na introdução, onde é apresentado um pequeno enquadramento do tema e motivação. Neste ponto, são apresentadas também os objetivos e resultados da dissertação; no capítulo 2 é desenvolvido um enquadramento teórico de todos os termos importantes para o desenvolvimento da dissertação, tais como: Sistemas de Informação na Saúde, Qualidade da Informação na Saúde, Sistema de Gestão da Qualidade, Descoberta de Conhecimento em Base de Dados, Sistemas de Apoio à Decisão, *Data Mining*, *Data Science*, *Business Intelligence* e Ontologias; no capítulo 3 são apresentadas abordagens metodológicas a utilizar no decorrer da dissertação, tais como: *Case Study*, *Design Science Research Methodology* e *Kimball Lifecycle.* Para além das abordagens metodológicas, é apresentada a metodologia utilizada, bem como ferramentas aplicadas e a orientação das tarefas; o capítulo 4 é referente a uma fase inicial da componente prática do projeto de dissertação, consistindo na aquisição do conhecimento onde são abordados temas como recolha e estudo dos dados, classificação do problema, desenho da solução e seleção e agregação dos dados; o capítulo 5 é inteiramente referente à componente prática do projeto de investigação, fazendo referência ao desenvolvimento da solução. Desta forma são abordados temas como a preparação dos dados, desenvolvimento do processo de Extract, Transform e Load (ETL) criação do cubo *Online Analytical Processing* (OLAP) introdução ao power BI, criação e análise de dashboards e elaboração de modelos de classificação e regressão.

Posteriormente à apresentação destes 5 grandes capítulos são apresentadas considerações finais do projeto, limitações e dificuldades do seu desenvolvimento e análise de riscos com a identificação das ações atenuantes para os riscos verificados.

2 Estado de Arte

Este capítulo reflete a revisão bibliográfica necessária para um bom desenvolvimento do projeto. O mesmo encontra-se dividido em diferentes secções, referentes a conceitos de relevância para um enquadramento ao tema, sendo eles estratégia de pesquisa, sistemas de informação na saúde, qualidade de informação na saúde, sistemas de gestão de qualidade, descoberta de conhecimento em base de dados, sistemas de apoio à decisão, *Data Mining* (DM), *Data Science* (DS), ontologias e trabalhos relacionados.

2.1 Estratégia de pesquisa

Relativamente ao enquadramento conceptual existiram algumas limitações a considerar, nomeadamente contextuais e temporais. Toda a documentação analisada e revista, corresponde a literatura científica disponível para o utilizador comum e para a comunidade académica da Universidade do Minho. No que diz respeito a plataformas de indexação de documentos destacam-se os seguintes: *Google Scholar*, *Repositorium* da Universidade do Minho, *Scopus*, *Web of Science, Elsevier's Science Direct* e *Google* (nomeadamente para pesquisas de informações fundamentadas não disponíveis em outras plataformas). Para uma melhor pesquisa, além dos documentos fornecidos pelo orientador, foi necessário definir alguns termos de pesquisa essenciais como: sistemas de informação na saúde, qualidade da informação na saúde, qualidade do serviço de reclamações, sistema de gestão de qualidade, importância da qualidade das reclamações na saúde, sistemas de apoio à decisão, Entidade Reguladora da Saúde (ERS), SNS (Serviço Nacional de Saúde), *data mining*, *data science*, *business intelligence*, *big data* e ontologias. No que diz respeito a limitações temporais, foram definidos limites para os documentos a analisar, nomeadamente com data igual ou superior a 2000, salvo raras exceções referentes a artigos ou autores relevantes para este projeto de dissertação. Para além disto, a revisão está limitada a documentos redigidos em língua Portuguesa e Inglesa.

2.2 Sistemas de Informação na Saúde

Ao longo dos anos, as organizações em geral sentem uma grande necessidade de se adaptar e acompanhar as transformações da sociedade. Esta necessidade de adaptação leva ao desenvolvimento de práticas de resposta, nomeadamente organização e disposição da informação sustentada por sistemas tecnológicos estruturados. Assim, todas estas mudanças e evoluções conduziram-nos para uma sociedade das tecnologias da informação, assumindo um papel preponderante em todos os setores e atividades da sociedade. É um facto que, hoje em dia, qualquer organização que queira entrar na

competitividade de mercado necessita de se adaptar facilmente a este. Desta forma, qualquer organização necessita de estar dotada com canais de comunicação eficazes e eficientes com o intuito de contornar inesperadas variações de mercado (Sandi, 2015).

O termo sistema de informação é considerado por muitos autores como complexo e amplo, não existindo uma definição concreta do termo, pois pode ser utilizado para designar coisas diferentes. Apesar da indefinição do termo, sabe-se que existem aspetos comuns nas várias utilizações, como por exemplo o facto de todos os sistemas de informação lidarem com informações, todos eles estão ligados a organizações ou ao trabalho realizado nas mesmas e todos estão ligados a tecnologias de informação por serem dependentes de computadores para a sua aplicação (Carvalho, 2000).

A mais recente definição referenciada por Carvalho (2000) no seu estudo remonta sistema de informação para um subsistema de um sistema organizacional, compreendendo a conceção, composição e funcionamento de aspetos como comunicação e informação de uma organização, descrevendo assim as orientações da informação e comunicação dentro dessa organização.

Atualmente, na área da saúde é indispensável a presença dos sistemas de informação, pois desempenham um papel fulcral na autonomização do utente relativamente à informação médica e de saúde (Espanha, 2010).

Segundo Sandi (2015) e tendo em conta a evolução tecnológica, torna-se cada vez mais impossível imaginar um serviço de cuidados de saúde sem que este utilize um sistema computorizado de informação pelas inúmeras vantagens que estão presentes na sua aplicação, como a notória melhoria do serviço ao utente e a facilidade de desempenho das tarefas dos funcionários.

A aplicação dos SI na saúde tem um propósito muito imperativo: contribuir para a qualidade do serviço, nomeadamente para o cuidado do utente de forma eficiente. Focando-se essencialmente no utente, com esta aplicação são expectáveis melhorias nos serviços médicos e de enfermagem prestados, nas tarefas administrativas e de gestão. Perante tais objetivos, torna-se evidente que o uso das tecnologias de informação traz diversas oportunidades para reduzir possíveis erros clínicos, tanto a nível económico como social. É de realçar a facilidade e velocidade de acesso a grandes quantidades de informação que os funcionários podem obter relativamente aos utentes, não esquecendo todas as outras atividades existentes num centro de cuidados de saúde (Sandi, 2015).

De acordo com Sandi (2015), atualmente existem alguns sistemas essenciais implementados em Portugal, nomeadamente: E-Agenda (marcação de consultas), E-SIGIC (consulta relativa a estado de cirurgia), RSE (registo de informação clinica) e WEBSIG (disponibiliza indicadores e metas do plano nacional de saúde).

Segundo os Serviços Partilhados do Ministério da Saúde (SPMS), o SClínico é um sistema de informação evolutivo desenvolvido pelos mesmos. Esta aplicação é centrada no doente e tem como principal objetivo ser única para todos os prestadores de cuidados de saúde. É uma aplicação que nasce da longa utilização de outras duas aplicações já existentes, o SAM (Sistema de Apoio ao Médico) e o SAPE (Sistema de Apoio à Prática de Enfermagem). O SClínico tem vindo a mostrar-se cada vez mais importante na área da saúde pois tem um papel fundamental na normalização da informação. O acesso à informação clínica do utente, a partilha e a utilização dos dados com profissionais de saúde de diversas áreas, são algumas das principais características desta aplicação. Assim, torna-se a atuação dos profissionais de saúde muito mais eficaz e eficiente, possibilitando um melhor apoio, assistência e acompanhamento ao utente (SPMS, 2017b).

Outra aplicação desenvolvida pela SPMS que tem tido muito sucesso é a Prescrição Eletrónica Médica (PEM). É uma aplicação utilizada em quase todo o SNS, sendo responsável por mais de 70% do total das prescrições registadas diariamente em Portugal. Tendo em vista o melhoramento das ferramentas de trabalho dos profissionais de saúde, a SPMS, pretende a curto prazo, que a PEM emita avisos aos médicos caso a sua prescrição não tenha sido levantada e outras situações de risco que o utente possa estar sujeito (SPMS, 2017a).

2.3 Qualidade da informação na Saúde

A qualidade da informação é um dos alicerces para a sobrevivência e maior competitividade das organizações. Ainda assim, a qualidade é cada vez mais exigida e igualmente importante no quotidiano da vida de qualquer ser humano.

O papel da qualidade da informação na saúde tem vindo a crescer exponencialmente ao longo dos anos. A qualidade da informação, bem como a disponibilização da qualidade cresce paralelamente aos avanços tecnológicos, obrigando assim, a adaptação das entidades prestadoras de cuidados de saúde. Desta forma, tornou-se essencial a disponibilização de informações na maior rede de comunicação, a internet. A Saúde que outrora era um assunto restrito, que poucas pessoas tinham acesso, a não ser que consultassem profissionais especializados de saúde, hoje torna-se disponível a qualquer pessoa. Um dos fatores negativos deste avanço é, efetivamente, a quantidade de informação que todos os dias surge na internet, pois muita desta informação pode não ser fidedigna induzindo em erro os leitores.

Desde a sua implementação, em meados da década de 1990, a qualidade da informação de saúde disponibilizada na internet despertou interesse nos profissionais de saúde, especialistas em informação

e consumidores. Com este rápido crescimento, surgiram várias iniciativas para avaliar a qualidade dos *websites* que disponibilizavam informações de saúde (Gagliardi & Jadad, 2002).

Segundo Berner (1999), é fundamental uma boa análise do conteúdo das informações consultadas por parte do utilizador de forma a confirmar a veracidade das mesmas, além disso, o utilizador deverá avaliar se a informação é efetivamente útil para ele, isto porque a maior parte das informações disponíveis são generalizadas, nomeadamente aspetos como sintomas de algum tipo de doenças que podem variar de utente para utente.

Atualmente os utilizadores podem confirmar a veracidade e atualização da informação através dos certificados presentes nos *websites*. Assim, a disponibilização de informações e intervenções na internet está a tornar-se cada vez mais importante na sociedade, sendo que é de extrema importância assegurar que as informações sejam imparciais, precisas, relevantes e oportunas para assegurar uma prestação de cuidados de saúde de qualidade (Berner, 1999).

No que toca à qualidade do registo de reclamações, é por muitos considerada uma componente essencial nos sistemas de saúde, tão importantes como as estratégias de promoção de cuidados. Um bom registo de reclamações pode contribuir eficazmente para o aprimoramento do sistema de saúde, possibilitando a identificação de falhas pontuais ou tendências. Além da melhoria da qualidade do serviço prestado, lidar efetivamente com as reclamações pode melhorar a comunicação interna entre profissional de saúde – utente, para não falar do aumento do nível de confiança e satisfação do utente (Vasconcelos Parra, 2014).

2.4 Sistema de Gestão da Qualidade

A qualidade da prestação de serviços deve ser medida de acordo com adequabilidade do serviço às especificações do cliente/utilizador final, ou seja, a qualidade deve medir-se de acordo com as respostas às necessidades do cliente.

A Norma Portuguesa ISO 9001 é uma norma de Sistemas de Gestão da Qualidade (SGQ) e está organizada de acordo com o ciclo Deming (Figura 1). Este ciclo tem como principal foco a melhoria contínua sendo uma ferramenta muito utilizada pelas organizações em todo o mundo. A fase inicial deste ciclo é o planeamento da ação, segue-se a execução da mesma, a monitorização e implementação de ações de melhoria contínua.

Figura 1 - Ciclo Deming (retirado de ("Ciclo PDCA," n.d.)*)*

Plan – Estabelecer os objetivos e os processos necessários para apresentar resultados de acordo com os requisitos do cliente;

Do – implementar os processos;

Check – monitorizar e medir processos e produto em comparação com políticas, objetivos e requisitos para o produto e reportar os resultados;

Act – empreender ações para melhorar continuamente o desempenho dos processos.

A gestão da Qualidade, de acordo com a ISO 9001 permite demonstrar o compromisso das organizações com a qualidade e a satisfação dos seus clientes/utentes. A ISO 9001 baseia-se em 8 princípios (SGS, 2017):

- Foco no cliente
- Liderança
- Envolvimento das pessoas
- Abordagem dos processos
- Abordagem da gestão como um sistema
- Melhoria contínua
- Abordagem factual
- Relações mutuamente benéficas com os fornecedores.

De acordo com estes princípios, as reclamações dos clientes são importantes para uma organização uma vez que estas dependem dos seus clientes, já por isso é que a ISO 9001 nos indica que devemos ter foco no cliente e envolvê-lo nos processos para que os produtos/serviços de uma organização respondam às necessidades dos clientes. Se isto acontecer, o número de reclamações irá diminuir levando a uma relação benéfica entre o cliente e a organização. A análise de reclamações ajuda a

organização na procura da melhoria contínua através da implementação de ações corretivas/preventivas, levando a uma melhoria da produtividade e redução dos custos de falhas/rejeições. O envolvimento das pessoas/clientes é fundamental na análise das reclamações.

2.5 Descoberta de Conhecimento em Base de Dados

As organizações estão constantemente a gerar informações e a armazená-las de uma forma descuidada e não organizada. No passado não era necessário olhar para esses dados para tomar uma boa posição face à competitividade do mercado. Hoje em dia, tal não acontece, pois as organizações sentem a necessidade de mudar e inovar para manter uma posição competitiva. Após alguns estudos, surge em 1989 o conceito de Descoberta de Conhecimento em Base de Dados (DCBD), que de alguma forma veio mostrar a importância da aquisição de conhecimento através das informações geradas (Fayyad, Piatetsky-Shapiro, & Smyth, 1996a). Assim, segundo Frawley, Piatetsky-Saphiro e Matheus (1992), a DCBD pode ser definida como um processo de extração de informação útil, conhecimento, a partir da leitura dos dados. Face aos casos de sucesso, as organizações perceberam que é fundamental ter em atenção os dados que armazenam, com o objetivo de melhorar o conhecimento através do recurso a técnicas de *Data Mining* (Frawley et al., 1992). Com o avanço tecnológico tornou-se possível elaborar um tratamento intrínseco e eficaz de todas as informações geradas, o que mudou muito a nível da gestão empresarial pois tornou-se mais percetível o processo de tomada de decisão. Atualmente, surge a necessidade emergente de recorrer a ferramentas tecnológicas para adquirir tal conhecimento, isto porque o volume de dados tem vindo a aumentar rapidamente ultrapassando a capacidade de processamento e análise humana (Oded & Rokach, 2010).

O processo de Descoberta de Conhecimento em Base de Dados além de ser iterativo, visto que pode existir retrocesso para etapas anteriores, é também interativo, visto que requer a participação do utilizador sempre que é necessária a tomada de decisão (Ramos & Santos, 2003). Assim, o DCBD é composto por cinco fases principais (Figura 2): Seleção, Pré-Processamento, Transformação, *Data Mining* e Avaliação/Interpetação.

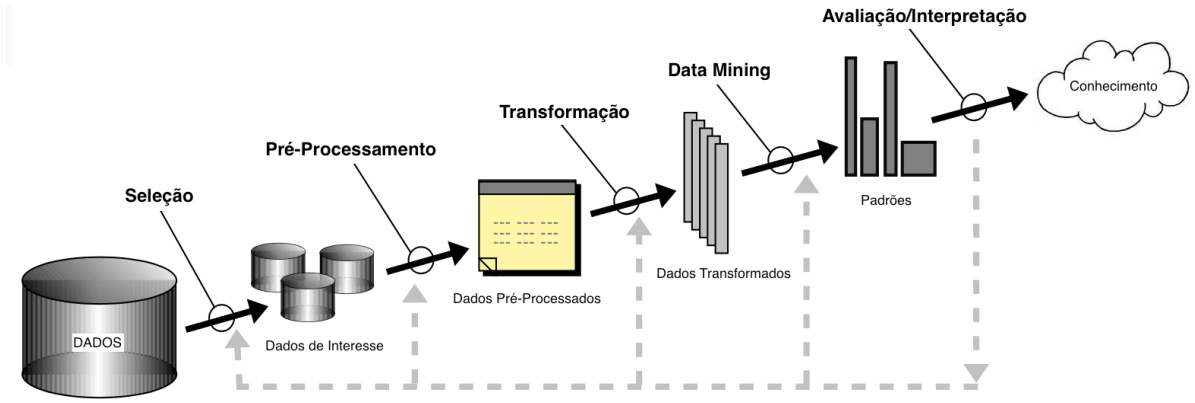

Figura 2 - Processo de DCBD (adaptado de (Fayyad et al., 1996a)*)*

Segundo os autores Oded e Rokach (2010), as cinco etapas principais são definidas como:

- **Seleção** – Definição dos objetivos e seleção dos dados a utilizar no processo de descoberta de conhecimento. Esta é uma etapa fundamental para uma boa solução final, isto porque a base de todo o processo para a construção dos modelos é definida neste passo sendo necessário uma boa definição dos atributos. Ainda nesta etapa, é importante descobrir todo o tipo de dados disponíveis para uma boa integração.
- **Pré-Processamento** – Nesta etapa pretende-se melhorar a fiabilidade dos dados, ou seja, os dados são sujeitos a filtragens, limpezas, alterações e remoções de forma a tornar os dados coerentes. Aqui são feitos vários estudos relativamente à fiabilidade dos atributos selecionados, sempre em busca dos melhores.
- **Transformação** – O objetivo principal desta fase é desenvolver os melhores modelos de dados. Se a qualidade dos dados a utilizar for boa, estes não sofrerão alterações. Por outro lado, se forem dados de fraca qualidade estes serão sujeitos a métodos/técnicas de transformação de forma a gerar dados úteis.
- ***Data Mining*** – Esta fase é fundamentalmente designada como a fase de escolhas consoante os objetivos definidos nas etapas anteriores. Aqui são selecionados métodos específicos bem como os padrões de pesquisa a utilizar de forma a atingir os melhores resultados possíveis.
- **Avaliação/Interpretação** – Esta etapa será aquela que irá delinear o fim do processo de DCBD ou o reajuste dos padrões de forma a repetir o processo. Nesta fase são avaliados e interpretados todos os padrões obtidos, tendo em atenção as metas definidas nas primeiras fases. A fiabilidade e a utilidade são sem dúvida regras fundamentais no processo de aceitação dos modelos.

Ao longo dos anos os conceitos DCBD e *Data Mining* foram sujeitos a definições muito parecidas, o que levou a perceções erradas dos termos. Na realidade são termos parecidos e um engloba o outro, mas

mesmo assim têm papeis distintos. Para Fayyad et al. (1996a), Descoberta de Conhecimento em Base de Dados refere-se ao processo geral de descoberta de conhecimentos úteis a partir de dados, e *Data Mining* refere-se a um processo particular neste processo. Os mesmos autores alertam para esta má perceção dos conceitos, além disso referem que uma aplicação exaustiva de métodos/técnicas de DM pode ser uma atividade perigosa, levando facilmente à descoberta de padrões sem sentido e inválidos. É de salientar que o DM depende fortemente da evolução da DCBD, que continua a evoluir a partir da interseção de campos de pesquisa, como a aprendizagem das máquinas, reconhecimento de padrões, base de dados e estatísticas.

Relativamente à descoberta de conhecimento existem algumas divergências relativas às opiniões desta tarefa. Segundo Fayyad et al. (1996a), a DCBD centra-se no processo global de descoberta de conhecimento a partir de dados, incluindo a forma como estes são acedidos e armazenados. Além disso, pode ser vista como uma atividade multidisciplinar que procura padrões compreensíveis que possam ser interpretados como conhecimentos úteis ou interessantes. Numa outra perspetiva, os autores Ramos e Santos (2003), defendem que os padrões extraídos durante o processo de descoberta de conhecimento não podem ser considerados conhecimento, uma vez que o conhecimento apenas pode residir na mente Humana em contínua ligação com a realidade interna e externa. Consideram ainda que estes padrões, em vez de conhecimento, devem ser apresentados como representações do conhecimento ou informação.

Atualmente, é fundamental um bom armazenamento de dados, de forma a suportar toda a informação gerada diariamente, sendo esta informação útil ou não. Este armazenamento de dados refere-se à tendência de armazenar e limpar os dados transacionais tornando-os acessíveis para análise e suporte à decisão (Fayyad et al., 1996a).

Após uma análise cuidada de várias abordagens da Descoberta de Conhecimento em Base de Dados, é possível salientar a definição mais conceituada e até mesmo considerada a mais correta para alguns autores. Os autores Fayyad et al. (1996b), têm várias publicações sobre esta temática, mas esta talvez seja a mais representativa que diz:

- "*A Descoberta de Conhecimento em Base de Dados é um processo não trivial para identificar padrões válidos, novos, potencialmente úteis e compreensíveis nos dados existentes.*"

2.6 Sistemas de Apoio à Decisão

Os primeiros Sistemas de Apoio à Decisão (SAD), do inglês *Decision Support Systems* (DSS), surgiram na década de 70 e desde logo têm vindo a surgir várias definições do conceito. Numa abordagem muito

simples foram definidos como sistemas computacionais interativos que ajudavam no processo de tomada de decisão. Mais tarde, em 1978, o mesmo autor designa SAD como sistemas de apoio à decisão que unem recursos intelectuais e computacionais para melhorar a capacidade de decisão, sendo estes interativos, adaptáveis e flexíveis (Turban, E. Aronson, & Liang, 2007).

Para Vercellis (2009), um sistema de apoio à decisão é uma aplicação interativa que combina dados e modelos matemáticos ajudando na resolução de problemas relativamente à gestão de organizações. *Business Intelligence* pode também ser considerado como um SAD, isto porque de certa forma influencia a transformação de informação em conhecimento útil para os gestores que irão realizar o processo de tomada de decisão.

Os SAD, têm estado em permanente evolução, desde a automatização de sistemas manuais caros até o fornecimento de valor organizacional estratégico. Neste seguimento surgem outros conceitos que melhoram a capacidade dos Sistemas de Apoio à Decisão, nomeadamente *Data Warehouse (DW)*[3] e *Data Mining*. De alguma forma, os SAD são agora uma parte vital de muitas organizações. A necessidade organizacional de combinar dados de múltiplos sistemas autónomos (por exemplo, financeiros, manufatura e distribuição) cresceu à medida que as organizações começaram a reconhecer o poder de combinar essas fontes de dados para realização de relatórios. Isso estimulou o crescimento dos sistemas de armazenamento de dados, por forma a conseguir armazenar um grande volume de dados para posterior análise (Nemati & D. Barko, 2010).

Para facilitar o processo de tomada de decisão, foram elaborados modelos estratégicos. Inicialmente o modelo mais comum, da autoria de Simon (1977), consistia em três fases fundamentais: Inteligência, Conceção e Escolha. Hoje em dia, o modelo reconhecido é o de Turban (2007) que sofreu alterações e foram acrescentadas mais duas etapas ficando assim constituído por cinco etapas, como é possível visualizar na Figura 3 : Inteligência, Conceção, Escolha, Implementação e Controlo.

[3] (*Data Warehouse)* – É utilizado para armazenar um conjunto de informações relativo às atividades de uma organização de uma forma consolidada. Possibilita a análise de grandes volumes de dados, sendo ainda definido como uma base de dados de grande dimensão que está organizada para dar suporte à tomada de decisões estratégicas da organização (Ribeiro, 2011).

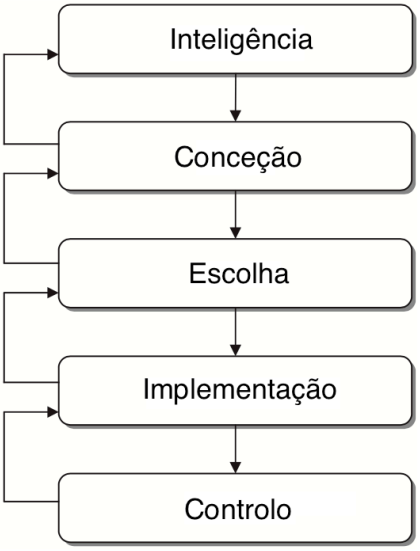

Figura 3 - Fases do processo de tomada de decisão (adaptado de (Vercellis, 2009)*)*

Segundo (Vercellis, 2009), as cinco fases do processo de tomada de decisão são definidas como:

- **Inteligência** – Nesta fase é identificado e definido explicitamente o problema que emerge no sistema em estudo. A análise do contexto e de toda a informação disponível é fundamental para que a pessoa responsável pela tomada de decisão identifique rapidamente o problema do sistema, dando assim instruções corretivas para o mesmo. A fase de Inteligência muitas das vezes resume-se a uma comparação entre o progresso atual das atividades com o plano de desenvolvimento original, sendo que é fundamental não confundir os sintomas com o problema em causa.
- **Conceção** – Nesta fase devem ser bem definidas as formas de resolver o problema identificado. Desta forma, devem ser desenvolvidas e planeadas as ações a tomar. Esta é uma etapa muito dependente da personalidade da pessoa responsável pela tomada de decisão, pois a experiência e a criatividade da pessoa desempenham um papel crítico. As soluções para resolver determinado problema advêm da capacidade de interpretação da pessoa, sendo diretamente dependentes da sua capacidade avaliativa. A pessoa fica assim responsável por encontrar várias alternativas para a resolução do problema.
- **Escolha** – Depois de identificar todas as alternativas para resolver o problema da organização em causa, torna-se necessário avaliá-las segundo critérios previamente definidos. Nesta fase podem ser seguidos métodos de otimização de forma a encontrar a solução mais correta, sendo que desempenham um papel muito valioso nesta fase de escolha.

- **Implementação** – Esta fase consiste na transformação da melhor alternativa em ações devidamente definidas num plano de implementação. Neste plano, devem constar ainda todas as funções e responsabilidades dos envolvidos no projeto, de forma a que todos os intervenientes percebam claramente o que têm de fazer.
- **Controlo** – Uma vez que foi feita uma boa implementação da solução para o problema, é necessário averiguar a integridade da mesma. É muito importante verificar se as expetativas iniciais foram satisfeitas e se as intenções da ação foram efetivamente conseguidas. A fase de controlo consiste nisto mesmo, em avaliar os resultados percebendo se todo este processo foi bem implementado. Posteriormente, todas as informações e experiências serão transferidas para o *Data Warehouse* ficando disponíveis durante processos de tomada de decisão subsequentes.

Hoje em dia, os SAD estão em constante crescimento devido ao facto de cada vez mais as decisões nas organizações serem tomadas por várias pessoas, em vez de uma só. Desta forma, a necessidade de reuniões e de trabalho de grupo aumenta na mesma proporção que aumenta a complexidade das tomadas de decisão, nomeadamente quando as decisões implicam a necessidade de se considerarem diferentes critérios (Marreiros, 2007).

Vários estudos têm sido feitos ao longo dos anos sobre esta temática, muitos deles complementa-se outros opõem-se em alguns aspetos. Um bom exemplo disso é o estudo de (Pomerol & Adam, 2004), defendendo que atualmente os SAD não estão suficientemente adaptados para apoiar a ação, definindo-os não só como deliberativos, mas também decisivos.

2.7 Data Mining

Data Mining é conhecido como um termo muito complexo que pode ser diretamente confundido com descoberta de conhecimento em base de dados ou simplesmente mineração de dados. O termo "mineração" advém dos tempos antigos aquando da exploração mineira, sendo a extração de dados uma analogia da extração de ouro das rochas (Han & Kamber, 1998).

Para (Han & Kamber, 1998) *Data Mining* é apenas um passo de todo o processo de descoberta de conhecimento, contudo aceitam ampliar esta visão definindo-o como um processo de descoberta de conhecimento de grandes quantidades de dados armazenados em bases de dados, DW ou outros repositórios de dados.

Da visão mais alargada de (Han & Kamber, 1998) surgiram vários estudos com o intuito de entender os tipos de padrões que poderiam ser extraídos. As próprias funcionalidades eram utilizadas para especificar

o tipo de padrões a ser encontrados, tendo então classificado DM em duas categorias: descrição e previsão. Definiram as tarefas de descrição como tarefas que caracterizam as propriedades gerais da base de dados, e tarefas de previsão como tarefas que identificam padrões dos dados de forma a realizar previsões.

O tema *Data Mining* tem sido alvo de muita atenção, talvez pela sua importância e influência nas organizações, surgindo vários estudos ao longo dos anos. Na verdade, segundo (Goebel & Gruenwald, 1999), DM consiste na extração eficiente de informação útil a partir de grandes volumes de dados, podendo afirmar-se que permite a descoberta de novos padrões, regularidades, factos e restrições facilitando a gestão de informação. Goebel e Gruenwald (1999) apresentam ideias diferentes relativamente à comparação dos termos DCBD e DM, sendo que, DCBD é utilizado para representar o processo de tornar os dados de baixo nível em conhecimento de alto nível, e DM é definido como a extração de padrões ou modelos de dados observados.

Segundo Fayyad et al. (1996a) ,*Data Mining* é talvez a etapa mais importante do processo de descoberta de conhecimento em base de dados que consiste na aplicação de algoritmos de análise e descoberta de dados produzindo padrões ou modelos. A definição mais emblemática e reconhecida é referente a este autor em que refere que DM consiste num processo não-trivial de identificar padrões válidos e potencialmente úteis e compreensíveis a partir de novos dados (Fayyad et al., 1996a).

Hoje em dia, é muito importante a presença de DM nas organizações, contudo é sempre necessário a presença de analistas humanos, que de alguma forma são sempre responsáveis por definir o valor do padrão encontrado. Os avanços na tecnologia são notáveis, mas infelizmente é imprescindível a presença de analistas capacitados que interajam com os sistemas de forma a conduzi-los para uma melhor extração de padrões úteis e relevantes (Navega, 2002).

É de notar que todos os estudos retratados anteriormente vão de encontro a perspetivas mais teóricas, sendo que é fundamental entender a diferença entre DCBD e DM. Numa perspetiva mais técnica, segundo Berry e Linoff (2004) os algoritmos de *DM* não foram inventados para generalizar, ou seja, é imprescendivel avaliar cada situação como única, avaliando as técnicas a ser utilizadas, a natureza dos dados e suas capacidades, bem como as preferências do utilizador.

Como anteriormente referido, os objetivos do DM podem ser divididos em duas grandes categorias, contudo podem ser divididas em subcategorias consoante o objetivo da sua utilização como é possível ver na Figura 4

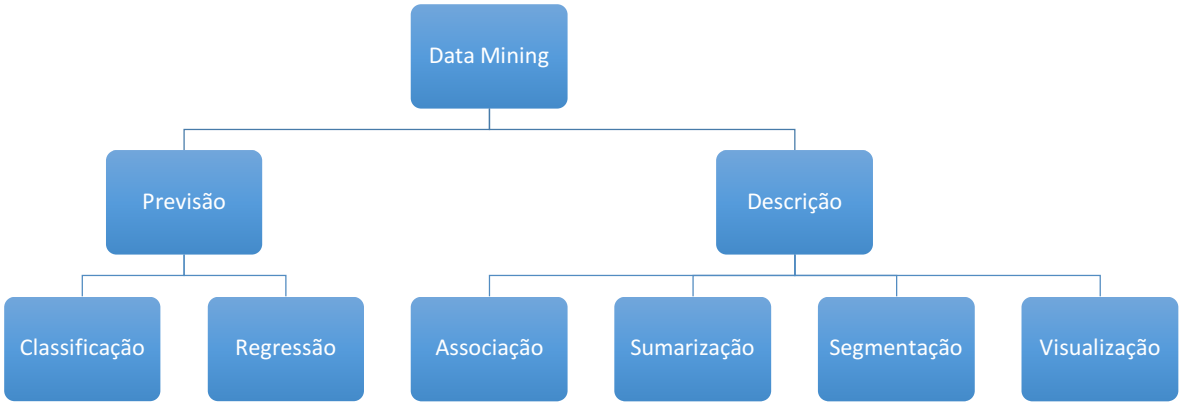

Figura 4 - Categorias e subcategorias de Data Mining (adaptado de (Pereira, 2005))

Como é possível visualizar na Figura 4, a categoria de previsão é constituída por classificação e regressão, a categoria de descrição é constituída por associação, sumarização, segmentação e visualização.

2.7.1 Classificação

A tarefa classificação é uma das tarefas mais comuns de *Data Mining*, estando diretamente ligada ao quotidiano de um ser humano, visto que o ser humano está constantemente a classificar e a categorizar objetos em classes. De uma forma geral, classificação consiste em atribuir classes a determinados objetos consoante as suas características, como por exemplo, atribuir raças a animais. No caso de um grande volume de dados, é possível descrever classificação como o processo de identificação de características de dados não classificados para que seja possível encontrar forma de os organizar em classes (Berry & Linoff, 2004; Pereira, 2005).

2.7.2 Regressão

A tarefa regressão é utilizada para definir um valor real a uma variável, como por exemplo prever a probabilidade de um individuo sobreviver a uma doença com base nos resultados de testes diagnósticos ou até mesmo prever o número de filhos de uma família (Fayyad et al., 1996a). Enquanto que a classificação lida com valores discretos, a regressão lida com valores contínuos (Pereira, 2005).

2.7.3 Associação

A tarefa associação é utilizada para determinar quais os dados que tendem em concorrer na mesma transação. O exemplo mais conhecido é a determinação dos produtos que costumam ser colocados juntos num carrinho de supermercado, análise de *market basket*. A associação destes produtos é

fundamental para os supermercados entenderem o consumidor de modo a planear a melhor forma de posicionar os artigos, isto é, adaptar a disposição das prateleiras de forma a colocar os produtos adquiridos na mesma compra próximos uns dos outros (Pereira, 2005). Esta tarefa é uma das mais conhecidas devido aos bons resultados obtidos, principalmente, no exemplo tratado.

2.7.4 Sumarização

A tarefa sumarização utiliza métodos para encontrar descrições detalhadas de um subconjunto de dados. Os métodos de sumarização mais evoluídos derivam de regras de resumo, técnicas de visualização e descobertas de relações entre variáveis. Estas técnicas de resumo são habitualmente aplicadas à análise de dados exploratórios e à geração automática de relatórios (Fayyad et al., 1996a).

2.7.5 Segmentação

A tarefa segmentação, do inglês *clustering*, é responsável na divisão heterogénea de uma população em vários subgrupos ou segmentos homogéneos. Neste processo não existem classes predefinidas sendo que os dados são agrupados conforme a sua semelhança (Fayyad et al., 1996a). Um bom exemplo de aplicação de segmentação é a associação de indivíduos com patologias iguais (Pereira, 2005).

2.7.6 Visualização

A tarefa de visualização trata das apresentações dos resultados de *Data Mining*, habitualmente através de gráficos e diagramas, permitindo uma boa representação de padrões e tendências. Esta tarefa é fundamental, sendo que deve ser bem fundamentada de forma a facilitar a compreensão da mesma (Pereira, 2005).

2.8 Data Science

Data Science (DS), do português ciência dos dados, é um conjunto de princípios fundamentais que apoiam e orientam a extração de informação e conhecimento dos dados. Um conceito muito próximo ao *Data Science* é o *Data Mining*. Estes princípios e técnicas podem ser aplicados em todas as áreas funcionais de negócio, mas são usualmente aplicados em tarefas de marketing, publicidade ou até mesmo em vendas. *Data Science* é também utilizado para monitorizar a relação com o cliente potenciando a sua satisfação. Pode ainda ser aplicado no setor financeiro em operações comerciais , gestão de clientes e em deteção de fraude. Muitas empresas destacam-se estrategicamente no mercado devido à aplicação de DS. Com uma eficaz aplicação de DS é possível a identificação de problemas no

negócio, considerado extração de conhecimento útil a partir dos dados em estudo. Esta análise só é possível a partir do pensamento analítico do cientista (Provost & Fawcett, 2013).

2.8.1 Business Intelligence

O sistema *Business Intelligence,* como o próprio nome indica, tende em ajudar e suportar novos conhecimentos, modificando e melhorando o processo de negócio de determinada organização, de forma a alcançar objetivos tornando-a mais competitiva no mercado em que se insere. Este sistema permite a combinação de dados operacionais com ferramentas analíticas, facilitando o processo de tomada de decisão dando enfase à qualidade e pontualidade do indicador de decisão. Através do BI torna-se mais fácil a compreensão das capacidades disponíveis na organização, bem como as orientações futuras no mercado, nas tecnologias e no ambiente em que compete (Negash, 2004).

Como consequência do volume de dados produzidos diariamente pelas organizações, surge a necessidade de encontrar algo que os suportasse para que fossem devidamente analisados e tratados. Este processo, dependendo do volume e da qualidade da informação disponível pode ser de curta ou longa duração. É estritamente necessário seguir um conjunto de processos para que o volume de dados esteja parametrizado conforme o que é pretendido no estudo. O aparecimento do *Data Warehouse* como repositório de grandes volumes de informações, os avanços nas ferramentas de análise, tratamento e limpeza de dados, o avanço da qualidade de processamento dos computadores são alguns dos aspetos que contribuíram efusivamente na melhoria do BI (Negash, 2004).

Em suma, o objetivo principal do BI passa por fornecer informação em tempo real, devidamente fundamentada e compreensível, melhorando o processo de tomada de decisão do gestor da organização. Segundo Negash (2004), os dados de análise produzidos pelas organizações podem ser estruturados e não-estruturados, posteriormente transformados em decisões, como é possível visualizar na Figura 5.

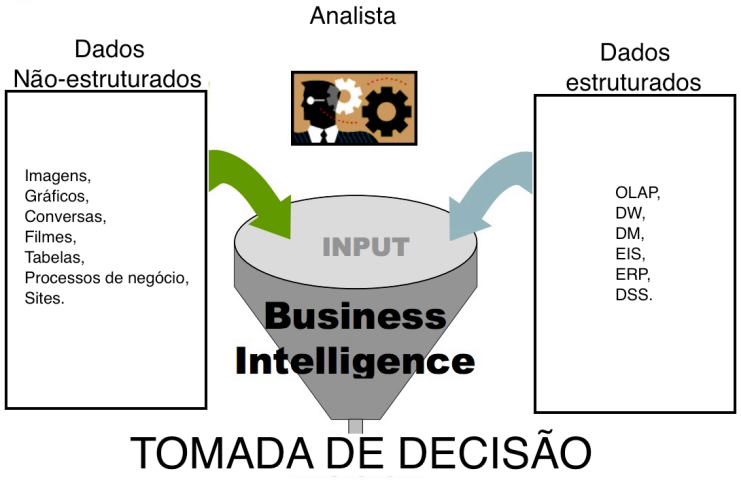

Figura 5 - Divisão de dados para sistemas de Business Intelligence (adaptado de (Negash, 2004)*)*

De uma forma muito simplista o processo de tratamento dos dados, desde a origem dos dados até a produção de indicadores de decisão, pode ser representado pela Figura 5. Assim, quando um analista de SI necessita de recorrer a sistemas de BI, o processo assemelha-se ao representado pela Figura 5. Assim, perante a origem dos dados, o analista identifica-os, analisa-os e trata-os de forma a que no final do processo seja possível identificar indicadores para a tomada de decisão.

2.8.2 Big Data

Atualmente é impensável não abordar o tema *Big Data* quando se fala em grandes volumes de dados, nomeadamente no que toca a produção de informação por parte das organizações. Neste contexto, é importante referir que *Big Data* não é nada mais que um termo abstrato, pois não é definido apenas por um grande volume de dados, mas sim como tudo o que o define (Chen et al., 2014).

A importância do *Big Data* é inquestionável, sendo que existe alguma ambiguidade no que toca à sua definição. Segundo Chen et al. (2014), de um modo geral, as definições baseiam-se em volumes de dados que não conseguem ser reconhecidos, adquiridos, geridos ou processados pelas tradicionais Tecnologias de Informação e software/hardware em tempo tolerável. Desta forma, esta definição remonta-nos para o foco da utilização do *Big Data*, visto que a integração de diferentes tipos de dados tem o objetivo a extração de informação útil para o negócio provocando vantagens competitivas de mercado.

Para Krishnan (2013), *Big Data* pode ser definido como quantidades de dados disponíveis com vários graus de complexidade e ambiguidade, sendo habitualmente gerados a diferentes velocidades. Perante os dados resultantes, as tecnologias, os métodos e algoritmos tradicionais deixaram de conseguir responder às necessidades. Assim, diretamente ligado a este conceito de *Big Data* está o modelo dos 3V's, nomeadamente volume, variedade e velocidade. O volume dos dados pode ser definido como a quantidade de dados gerada continuamente com diferentes tipos e tamanhos. Relativamente à variedade de dados, esta pode ser definida como os múltiplos formatos de dados possíveis tais como: estruturados, semiestruturados ou não estruturados. Por fim, a velocidade dos dados consiste no tempo de processamento de forma a atingir os resultados esperados.

Atualmente, o modelo dos 3V's pode agregar mais três características complementares como ambiguidade, viscosidade e viralidade, tornando este modelo mais completo. A ambiguidade é criada pelo volume e variedade dos dados associados à indisponibilidade dos metadados. A viscosidade é caracterizada pela resistência do volume de dados, resultante das regras de negócio ou limitações tecnológicas. A viralidade é definida pelo tempo que os dados demoram a propagar-se na rede. Desta

forma, através do cruzamento destas características é possível atenuar a complexidade associada ao processamento de *Big Data*. (Krishnan, 2013)

2.9 Ontologias

Com o aumento exponencial dos dados tem vindo a ser acrescentada mais importância e responsabilidade a técnicas de organização da informação. Atualmente, neste contexto, tem sido notória a atenção dada à utilização de ontologias na organização dos dados. Uma ontologia pode ter um significado muito amplo e tende a variar conforme o objetivo da sua utilização. Numa perspetiva mais técnica, uma ontologia é definida como um conjunto de termos, com ordenação hierárquica, descrevendo um domínio específico para ser utilizado em bases de conhecimento (M. Almeida & Bax, 2003).

Segundo A.Moreira (2002), as ontologias são utilizadas para desenvolvimento de sistemas computacionais, elas podem também ser utilizadas especificamente no desenvolvimento de sistemas de informação mais flexíveis e fáceis de utilizar. Por outro lado, as teorias geradas pelos estudos sobre classificação da ciência da informação podem ajudar na criação de metodologias para o desenvolvimento de ontologias mais robustas.

Em suma, num modo geral, o termo ontologia denota o conjunto de conceitos e relações comuns a um determinado domínio. Esses termos podem ser expressos em linguagens naturais ou formais, como uma teoria. No entanto, uma ontologia pode possuir diversas representações e uma teoria pode expressar diversas ontologias (A. Moreira, 2002).

Na área das ciências computacionais é cada vez mais importante uma boa organização de toda a informação, sendo que as reclamações na área da saúde não são exceção. É fundamental melhorar, a todos os aspetos, a qualidade da informação, de forma a que a perceção de determinado incidente seja mais eficaz. No que toca à saúde é imperativa uma boa organização da informação, sendo que a criação de ontologias com determinados relacionamentos e dependências pode, e muito, influenciar uma melhor gestão a este nível.

2.10 Trabalho relacionado

Atualmente a saúde tem um papel fundamental na sociedade e está claramente presente no dia-a-dia de todas as pessoas. Estas mesmas pessoas já não recorrem aos serviços prestadores de cuidados de saúde apenas em último recurso, isto é, quando necessitam efetivamente de algum serviço, recorrem sim regularmente pois preocupam-se cada vez mais com o seu bem-estar. Com isto, é expectável que os serviços sejam melhorados, acabando por ser exigido cada vez mais dos serviços.

No seguimento surge a parte das reclamações na área hospitalar que é cada vez mais recorrente devido a insatisfação pelo serviço prestado. Deste modo, esta ainda é uma área pouco abordada, contudo a área da gestão das reclamações tem sido bastante cobiçada e estudada através de sistemas desenvolvidos e aplicados.

Como foi referido, no que toca à qualidade e auditoria das reclamação em si não existem muitos estudos, mas a dissertação de mestrado de André Agostinho Granja da Silva Oliveira (2015) com o tema – Apoio à Decisão na Análise Inteligente de Reclamações remonta para a visualização da informação contida nas reclamações, exploração e criação de modelos de classificação e sugestão automáticos de reclamações recolhidas em unidades prestadoras de cuidados de saúde.

Relativamente a estudos referentes ao uso de *Data Mining* na saúde em Portugal existem alguns que devem ser mencionados, como a dissertação de mestrado de Daniela da Silva Alves (2015) com o tema – Saúde em Portugal: Estudo das Urgências Hospitalares através do *Data Mining*, abordando o estado do serviço de urgências quando solicitado por utentes que não necessitam do mesmo, limitando por vezes a qualidade do serviço bem como o tempo de resposta. Um outro estudo a ser considerado é relativo ao tema – *Data Mining* e Sistemas de Apoio à Decisão em Aplicações Clínicas e Qualidade de vida, abordando a obtenção de conhecimento a nível de padrões comportamentais de utentes crónicos, com o objetivo de potenciar o processo de tomada de decisão por parte das equipas médicas especializadas (M. Ferreira, Reis, Gonçalves, & Rocha, 2015).

3 Abordagem Metodológica

Este capítulo reflete as aboordagens metodológicas e as ferramentas utilizadas no desenvolvimento do projeto. O mesmo encontra-se dividido em secções, sendo elas Case Study, Design Science Research, Kimball Lifecycle, metodologia utilizada, ferramentas utilizadas e orientação de tarefas.

Relativamente às abordagens metodológicas a utilizar é importante referir que neste projeto de dissertação foi utilizado um caso de estudo proveniente da Entidade Reguladora da Saúde (ERS), tendo por base um estudo realizado sobre as reclamações no projeto de investigação do André Agostinho Granja da Silva Oliveira (Oliveira, 2015) . Deste modo, para uma melhor resposta a este caso de estudo foi utilizada uma conjugação de metodologias, nomeadamente o *Case Study*, *Design Science Research Methodology* e o *Kimball Lifecycle*. Com esta conjugação foram desenvolvidos artefactos e experiências capazes de responder às necessidades do estudo em causa. Ainda neste contexto foram abordadas investigações relativas à metodologia *Crisp-DM*, que neste projeto não foi aplicada devido a atrasos verificados provocando a não realização do *Data Mining* (DM).

3.1 Case Study

A abordagem metodológica *Case Study*, do português estudo de caso, é considerada por muitos investigadores um método de pesquisa muito robusto, quando é particularmente necessário uma investigação mais aprofundada. É uma ferramenta muito reconhecida em casos de estudo relacionados com ciências socias, embora também seja muito utilizada em questões de educação, sociologia e comunidade. O método estudo de caso, acaba por surgir com a necessidade de ultrapassar as limitações dos métodos quantitativos tradicionais, no que toca a explicações aprofundadas de problemas comportamentais e sociais. Através deste método, é permitido a um investigador ir além de meros resultados estatísticos, passando a compreender as condições comportamentais com base em perspetivas, significando que um mesmo caso poderá apresentar várias perspetivas consoante o número de investigadores. Em suma, o estudo de caso conjuga métodos qualitativos e quantitativos, melhorando a perceção tanto do processo como do resultado de um fenómeno, através da observação completa dos casos de investigação (Zainal, 2007).

Segundo Johansson (2003), este método, quando aplicado apenas a um único caso, poderá apresentar limitações nas suas conclusões, isto porque não podem ser apresentadas de forma generalizada. Assim, surge a necessidade de conjugar vários métodos aquando da utilização do método estudo de caso. A Figura 6 representa as várias estratégias que podem influenciar a metodologia estudo de caso. Este método permite a combinação de seis grandes estratégias: argumentação lógica, interpretação histórica,

qualitativo, correlacional, experimental, simulação. Segundo Johansson (2003), não existe uma definição específica para cada uma destas combinações, no entanto, existe uma lógica que suporta o modelo. No momento em que o investigador inicia uma pesquisa interpretativa e qualitativa deve ter uma abordagem comum ao tema, mas com diferentes perspetivas temporais. Por outro lado, a pesquisa correlacional é partilhada com a pesquisa qualitativa e dependente da quantidade dos dados como a pesquisa experimental. Juntamente com a pesquisa experimental a simulação requer controlo e manipulação. A argumentação lógica depende da análise temporal, sendo fundamental para uma boa interpretação histórica da pesquisa efetuada (Johansson, 2003).

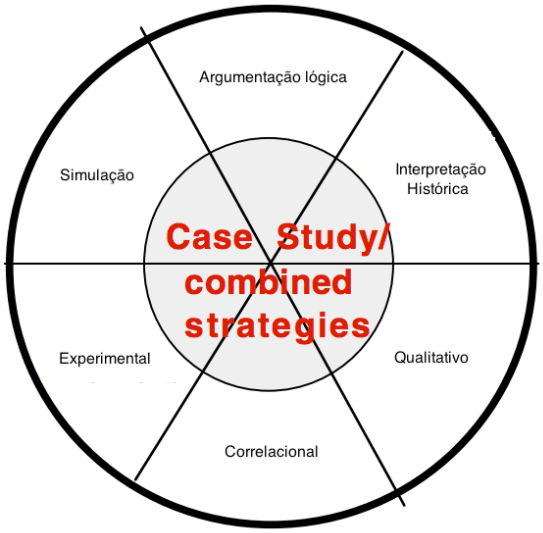

Figura 6 - Conjugação de vários métodos (adaptado de (Johansson, 2003)*)*

Segundo Zainal (2007), a conceção cuidadosa de um estudo de caso é portanto muito importante, pois este método utilizado através de entrevistas ou de diários, deve ser capaz de provar que:

- É o único método viável para obter dados implícitos e explícitos do sujeito;
- É apropriado para a questão de pesquisa;
- Segue o conjunto de procedimentos com aplicação adequada;
- As convenções científicas utilizadas nas ciências sociais são rigorosamente seguidas;
- O conjunto de provas, quantitativas e qualitativas, são sistematicamente gravadas e arquivadas, particularmente quando as principais fontes são entrevistas e observações diretas do investigador;
- O estudo de caso está ligado a um quadro teórico.

Segundo Zainal (2007), existe uma série de vantagens e desvantagens no uso de estudos de caso, sendo que é importante não confundir estudos de caso com pesquisas qualitativas e em alguns casos, estudos de caso podem ser baseados inteiramente em provas quantitativas.

Vantagens:

- Distingue-se na forma de compreensão dos dados, visto que o investigador tem que perceber o ambiente em que o problema ocorre;
- Variações em termos de abordagens intrínsecas, instrumentais e coletivas para estudos de caso, permitem análises qualitativas e quantitativas dos dados;
- Evidências de respostas numéricas e categóricas de indivíduos;
- Análises qualitativas produzidas, não só ajudam a explorar ou a descrever os dados em ambiente real, como também ajudam a explicar as complexidades de situações reais que não podem ser observadas através de pesquisas experimentais.

Desvantagens:

- Estudos de caso são sistematicamente acusados de falta de rigor;
- Fornecem poucas bases para a generalização científica;
- Demasiado longos, difíceis de conduzir e uma enorme quantidade de documentação.

No desenvolvimento deste projeto de dissertação, a metodologia científica *case study* acompanhou todas as fases do mesmo, sendo que a sua utilização foi fundamental no que toca à conjugação de variados métodos qualitativos e quantitativos, melhorando a perceção tanto dos processos como dos resultados alcançados, através da observação completa de todos os casos de investigação. Assim, o caso de estudo aplicado é proveniente das reclamações em papel e online inseridas pelos utentes de serviços prestadores de cuidados de saúde, tendo por base um estudo realizado no passado pelo André Agostinho Granja da Silva Oliveira (2015).

3.2 Design Science Research Methodology

Segundo os autores Peffers, Tuunanen, Rothenberger e Chatterjee (2007), uma metodologia é um sistema de princípios, práticas e procedimentos que são aplicados a um determinado ramo específico do conhecimento. No estudo realizado, entende-se que a metodologia *Design Science Research* é utilizada para facilitar a pesquisa a profissionais de sistemas de informação, sendo que através desta metodologia é possível produzir e apresentar documentos de alta qualidade. O modelo apresentado na Figura 7, é uma junção de vários modelos apresentados por outros autores, ou seja, em vez de

apresentarem uma ideia diferente, estes autores decidiram criar uma metodologia que serviria de estrutura genérica aceite e contemplada por todos.

A metodologia de investigação científica é constituída por seis etapas fundamentais: Identificação problema e motivação, Definição dos objetivos da solução, *Design* e conceção, Demonstração, Avaliação e Comunicação, não sendo necessário seguir uma ordem sequencial de etapas.

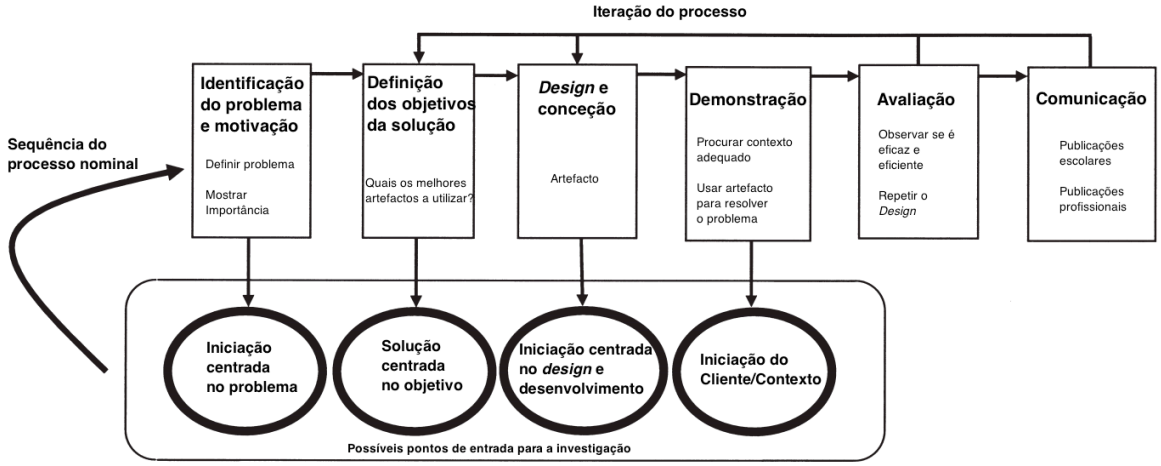

Figura 7 - Fases da metodologia Design Science Research (adaptado de (Peffers et al., 2007)*)*

3.2.1 Identificação do problema e motivação

Definição da pesquisa específica para o problema em causa, seguida da justificação da solução com auxílio a artefactos. O processo de justificação da solução motiva a pessoa que está a desenvolver o estudo e o leitor. Para o desenvolvimento desta etapa é necessário o conhecimento do estado do problema e da importância da sua solução (Peffers et al., 2007).

Este ponto foi utilizado como referência para a identificação do problema estudado neste projeto de dissertação, bem como o desenvolvimento da motivação para a realização do mesmo.

3.2.2 Definição dos objetivos da solução

Definição dos objetivos da solução com base na definição do problema e do conhecimento, estes objetivos podem ser quantitativos se a solução desejável for melhor que a atual, e qualitativos se o artefacto esperado influenciar a solução para problemas até então não encontrados. É necessário o conhecimento do estado do problema e da solução atual para melhorar a sua eficácia (Peffers et al., 2007).

Este ponto foi utilizado como referência para a identificação e desenvolvimento dos objetivos deste projeto, de forma a atingir uma solução desejável.

3.2.3 *Design* e conceção

Criação de artefactos, podendo ser modelos, métodos ou instâncias. Teoricamente, um artefacto de pesquisa pode ser qualquer objeto projetado que contribui na pesquisa incorporada no *design*. Esta tarefa permite determinar as funcionalidades desejadas do artefacto bem como a sua arquitetura, sendo posteriormente possível a sua criação. É necessário adquirir conhecimento teórico para determinar a solução (Peffers et al., 2007).

Este ponto foi tomado como referência para o estudo das funcionalidades e o desenvolvimento da arquitetura tendo como fim alcançar os objetivos traçados.

3.2.4 Demonstração

Demonstração do uso do artefacto para solucionar uma ou mais instâncias do problema. Poderá ser necessário realizar experiências, simulações, casos de estudo ou outra atividade apropriada. É necessário entender como o artefacto poderá solucionar o problema (Peffers et al., 2007).

Este ponto foi tomado como referência para o desenvolvimento de testes de modo a garantir que todo o processo até então criado é o melhor para responder ao problema em causa.

3.2.5 Avaliação

Observação da capacidade do artefacto solucionar o problema. Esta tarefa envolve um processo de comparação dos objetivos da solução com os resultados reais após a utilização do artefacto na demonstração. São necessários conhecimentos de métricas e técnicas relevantes, dependendo da natureza do problema. A natureza da pesquisa pode ser essencial para determinar a viabilidade da iteração (Peffers et al., 2007).

Este ponto foi tomado como referência para determinar se os resultados atingidos vão, efetivamente, de encontro com os resultados expectáveis e de alguma forma perceber se estes se adequam à realidade.

3.2.6 Comunicação

Comunicação do problema, do artefacto bem como dos resultados obtidos ao público adequado. Normalmente em publicações académicas, poderá ser feita a exposição numa estrutura de papel ou numa estrutura de pesquisa experimental. É necessário adquirir conhecimento linguístico, cultura disciplinar (Peffers et al., 2007).

Este ponto acompanhou grande parte do projeto de dissertação, pois foi tomado como referência em tudo o que toca à escrita dos documentos científicos para exposição de resultados.

3.3 Kimball Lifecycle

Segundo Ross (2009), a metodologia do ciclo de vida *Kimball* foi concebida durante meados da década de 1980 por membros do grupo *Kimball* e membros da *Metaphor Cumputer Systems*. Desde a sua conceção tem vindo a ser cada vez mais utilizada em projetos de *Business Intelligence (*BI*)* e *Big Data (*BD*)* em praticamente todos os setores. Esta metodologia baseia-se em três princípios fundamentais:

- Foco na adição de valor de negócios em toda a organização;
- Dimensionar os dados que são entregues ao negócio;
- Desenvolver iterativamente o ambiente BD/BI em incrementos de ciclo de vida.

De forma a garantir um projeto bem-sucedido, é fundamental seguir as tarefas de alto nível ilustradas na Figura 8.

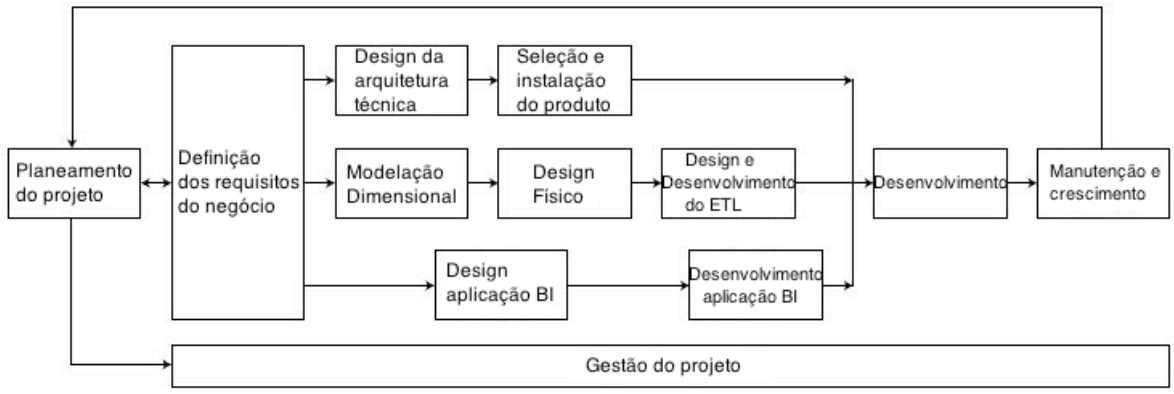

Figura 8 - Ciclo de vida Kimball (adaptado de (Ross, 2009))

3.3.1 Planeamento do Projeto

Esta fase centra-se no planeamento e âmbito do projeto, sendo tarefas contínuas mantendo todas as atividades sob controlo (Ross, 2009).

Este ponto é utilizado como referência ao longo do projeto garantindo o cumprimento de todas as tarefas planeadas atempadamente.

3.3.2 Definição dos requisitos do negócio

Esta fase é importante para definir claramente os objetivos do projeto. É neste ponto que todos os requisitos são analisados aprofundadamente, garantindo que os requisitos escolhidos são claramente os melhores para o desenvolvimento do projeto (Ross, 2009).

Este ponto é utilizado como referência na identificação dos requisitos e objetivos do problema do projeto.

3.3.3 Design e seleção da arquitetura

Em projetos de BD/BI é fundamental a integração de várias tecnologias de armazenamento de dados. Desta forma, esta tarefa está associada ao design da arquitetura do sistema para satisfazer as necessidades do negócio (Ross, 2009).

Este ponto foi tomado como referência para o estudo das funcionalidades e o desenvolvimento da arquitetura com o fim de alcançar os objetivos traçados.

3.3.4 Modelação dimensional e desenvolvimento do ETL

A modelagem dimensional é onde os dados são divididos em factos de medição ou dimensões descritivas. Modelos relacionais podem ser instanciados em bases de dados relacionais, como modelos em estrela, modelos multidimensionais e cubos *Online Analytical Processing* (OLAP). É fundamental garantir facilidade de uso da perspetiva de utilizador e desempenho rápido das consultas. No processo de transformação dos dados é fundamental seguir alguns critérios como: extrair os dados da fonte; realizar transformações de limpeza e conformidade; garantir uma boa gestão dos processos em ambiente de *Extract, Transform e Load* (ETL) (Ross, 2009).

Este ponto foi tomado como referência em todas as atividades presentes no tratamento dos dados.

3.3.5 Design e desenvolvimento de aplicações BI

Nesta fase são identificadas e construídas uma basta gama de aplicações BI incluindo relatórios padronizados, consultas parametrizadas, *dashboards*, modelos analíticos e aplicações de tratamento de dados (Ross, 2009).

Este ponto foi tomado como referência no processo de análise dos dados e na apresentação dos resultados garantindo uma boa perceção dos mesmos.

3.3.6 Implementação e manutenção

O ciclo de vida *Kimball* converge na implementação, reunindo todas as aplicações de tecnologia bem como todo o tratamento de dados. A fase de manutenção é muito importante em todo este processo, visto que é um processo de longo prazo. Em todo o projeto é importante estar a par de todas as limitações e riscos, pois o grande objetivo, independentemente da organização em causa, é a aceitação de negócios para apoiar o processo de tomada de decisão (Ross, 2009).

Este ponto foi tomado como referência no processo de desenvolvimento dos melhores indicadores para a tomada de decisão.

3.4 Metodologia utilizada

Para o desenvolvimento deste projeto de dissertação, foi decidido em conjunto com os orientadores que a metodologia ideal seria a combinação de três metodologias apresentadas anteriormente (*Case Study*, *DSRM* e *Kimball Lifecycle*). Assim foi desenvolvida a Tabela 1 que representa a conjugação das três metodologias com o acréscimo do *Crisp-DM* a aplicar no desenvolvimento de *Data Mining*. É possível identificar as tarefas das várias metodologias que terão influência no desenvolvimento de cada etapa do projeto.

Tabela 1 - Cruzamento de metodologias

	Case Study	DSRM	Crisp-DM	Kimball Lifecycle
Etapa 1	Argumentação lógica; Interpretação histórica	Identificação do problema e motivação; Definição de objetivos;	Compreensão do negócio; Compreensão dos dados	Planeamento do projeto; Definição de requisitos do negócio
Etapa 2	Qualitativo; Correlacional	Design e conceção;	Preparação dos dados; Modelação	Design e seleção da arquitetura; Desenvolvimento do ETL e modelação dimensional; Design e desenvolvimento de aplicações BI
Etapa 3	Experimental			
Etapa 4	Simulação	Demonstração; Comunicação;	Avaliação	
Etapa 5		Avaliação	Implementação	Implementação e manutenção

3.5 Ferramentas Utilizadas

No decorrer deste projeto de investigação, nas variadas etapas, foi necessário o recurso a diversas ferramentas. Assim, na Tabela 2 estão identificadas as ferramentas utilizadas bem como a respetiva descrição.

Tabela 2 - Lista de ferramentas utilizadas

Ferramenta	Descrição
Microsoft Word 2016	O *Microsoft Word* 2016, disponibilizado pela *Microsoft* no pacote Office, permite facilmente a criação de documentos de texto. Esta ferramenta foi utilizada para a formatação de texto e elaboração de documentação.
Microsoft Excel 2016	O *Microsoft Excel* 2016, disponibilizado pela *Microsoft* no pacote Office, permite a análise e exploração de dados. Esta ferramenta foi utilizada essencialmente na exploração dos dados.
MySQL Workbench	O *MySQL Workbench* permite o desenho e criação de base de dados. Esta ferramenta foi utilizada para o primeiro armazenamento dos dados, bem como para os desenhos dos modelos relacionais.
Microsoft SQL Server 2012	O *Microsoft SQL Server* 2012 permite a criação e gestão de base de dados. Esta ferramenta foi utilizada para armazenar e gerir os dados escolhidos.
Microsoft Visual Studio 2010	O *Microsoft Visual Studio* 2010 através do pacote de Business Intelligence permite o desenvolvimento do processo ETL, *Reports* e *Dashboards*. Esta ferramenta foi utilizada para o desenvolvimento de todo o processo ETL bem como o desenvolvimento do cubo, recorrendo aos módulos *Integration Services* e *Analysis Services*, respetivamente.
Microsoft Power BI	O *Microsoft Power BI* permite efetuar análises do negócio através de *Dashboards*, com base em determinados dados. Esta ferramenta foi utilizada para criação de relatórios e consequentes *Dashboards* dos dados tratados.

3.6 Orientação de tarefas

Esta fase do projeto de investigação centra-se essencialmente na organização da sua parte mais prática, isto é, com base na abordagem metodológica desenvolvida anteriormente, onde são agregadas várias fases das quatro metodologias estudadas, evidenciar cada etapa às tarefas correspondentes. Assim, como é visível na Tabela 3, esta parte mais prática do projeto dividiu-se em duas grandes componentes: aquisição do conhecimento e desenvolvimento da solução. A componente de aquisição do conhecimento contempla várias tarefas correspondentes à primeira etapa. Relativamente à componente de desenvolvimento da solução, mais abrangente, contempla várias tarefas correspondentes às etapas dois, três e quatro.

Tabela 3 - Lista de tarefas correspondentes a cada etapa do projeto

	Aquisição do conhecimento	Desenvolvimento da solução
Etapa 1	Recolha e estudo dos dados Classificação do problema Desenho e solução Seleção e agregação de dados de investigação	
Etapa 2 **Etapa 3** **Etapa 4**		Preparação dos dados Processo *ETL* Criação do cubo Introdução ao *Power BI* Processo de criação de *dashboards* Análise dos *dashboards* desenvolvidos Elaboração de modelos de regressão e classificação

4 AQUISIÇÃO DO CONHECIMENTO

Este capítulo reflete o primeiro contacto com o *dataset*, estando dividido em três secções, entre elas a recolha e estudo dos dados, a classificação do problema e o desenho da solução. Na primeira secção é possível observar o resultado das primeiras análises desenvolvidas ao *dataset* fornecido. Na secção seguinte é apresentada a classificação do problema central em estudo. Por último, na secção do desenho e solução são apresentadas as alterações príncipais a que os dados foram sujeitos.

4.1 Recolha e estudo dos dados

Para o desenvolvimento deste projeto de dissertação era esperada a disponibilização de dados de investigação atualizados por parte da Entidade Reguladora da Saúde (ERS). Não sendo possível e numa perspetiva de solução para o tema, foram disponibilizados dados pelos orientadores, que já teriam sido objeto de estudo pelo André Agostinho Granja da Silva Oliveira (2015) na sua dissertação de mestrado. Como referido na secção 2.10 este estudo centra-se na visualização da informação contida nas reclamações. Sendo que o único aspeto comum neste projeto de investigação com o do André Agostinho Granja da Silva Oliveira (2015) são os dados e a tipologia dos mesmos.

Assim, na sequência deste projeto, os dados disponibilizados são relativos a reclamações online e reclamações em papel.

Os dados disponibilizados são resultantes das reclamações efetuadas por utilizadores de entidades prestadoras de cuidados de saúde, que de alguma forma sentiram a necessidade de se manifestarem. Ao contrário do que o próprio nome indica, estes utilizadores não recorrem apenas às reclamações para expressar opiniões negativas, servindo também como meio de comunicação para partilha de opiniões positivas à entidade responsável por receção e tratamento destas opiniões, a ERS. Estes dados, resultam disso mesmo, da agregação de todas as opiniões partilhadas por estes utilizadores através de livros de reclamações (reclamações em papel) e do portal de reclamações (reclamações online). A ERS, como qualquer outra entidade responsável por agregação de informação é totalmente imparcial à informação contida nas reclamações, sendo que é perfeitamente normal encontrar informação inútil sem possibilidade de análise. O contrário também é visível neste tipo de reclamações, ou seja, é possível encontrar críticas construtivas, testemunhos positivos e até mesmo agradecimentos relativos às boas práticas desenvolvidas pelas entidades prestadoras de cuidados de saúde.

Os dados recebidos foram sujeitos a várias análises, até porque não eram de fácil compreensão, uma vez que as reclamações/sugestões em si não eram objetivas, não havendo descrições do *dataset*. É importante salientar que todo o projeto de investigação, inclusive o processo de análise e tratamento dos

dados teve um cariz profissional, através do anonimato foi possível garantir total confidencialidade dos mesmos. Desde logo, numa análise superficial, foram identificados problemas de má estruturação dos dados. Nesta primeira análise resultou a seguinte estrutura:

- ers_reclamacoesonlineestados (Tabela 4)
- ers_reclamacoes_tipos_diligencias (Tabela 5)
- ers_ac_valencias (Tabela 6)
- ers_tipificacao (Tabela 7)
- ers_reclamacoesonline (Tabela 8)
- ers_reclamacoes (Tabela 9)

Tabela 4 - Estrutura da tabela ers_reclamacoesonlineestados

Tabela	Descrição	Nº Registos	Campos	Tipo	Erros Comuns
ers_reclamacoesonlineestados	Estados possíveis da reclamação	6	Id descricao	*INT* *VARCHAR(20)*	

Tabela 5 - Estrutura da tabela ers_reclamacoes_tipos_diligencias

Tabela	Descrição	Nº Registos	Campos	Tipo	Erros Comuns
ers_reclamacoes_tipos_diligencias	Tipos de diligências	5	Id descricao	INT VARCHAR(20)	

Tabela 6 - Estrutura da tabela ers_ac_valencias

Tabela	Descrição	Nº Registos	Campos	Tipo	Erros Comuns
ers_ac_valencias	Valências	44	Id nome valido	INT VARCHAR(40) VARCHAR(5)	

Tabela 7 - Estrutura da tabela ers_tipificacao

Tabela	Descrição	Nº Registos	Campos	Tipo	Erros Comuns
ers_tipificacao	Tipificação da reclamação	496	id descricao valido pailD nivel propoeArquivamentoREC	INT VARCHAR(100) VARCHAR(5) INT INT VARCHAR(5)	

Tabela 8 - Estrutura da tabela ers_reclamacoesonline

Tabela	Descrição	Nº Registos	Campos	Tipo	Erros Comuns
ers_reclamacoesonline	Dados de reclamações online	1816	descricao datacriacao dataalteracao estadoid reclamacaoid encaminhadopara ano numero tipificacaoid tipologia razaoarquivamentoliminar razaoignorada	*VARCHAR(5000)* *VARCHAR(20)* *VARCHAR(20)* *INT* *INT* *VARCHAR(20)* *INT* *INT* *INT* *VARCHAR(20)* *VARCHAR(20)* *VARCHAR(20)*	Ausência de chave primária. Existência de muitos *NULL* – Identificar o sentido da sua utilização.

Tabela 9 - Estrutura da tabela ers_reclamacoes

Tabela	Descrição	Nº Registos	Campos	Tipo	Erros Comuns
ers_reclamacoes	Dados de reclamações escritas	12332	n_folha_reclamacao data_entrada sintese_reclamacao proposta_actuacao valencia apreciaCAo_clinica_id assunto_visado_outro	*VARCHAR(20)* *VARCHAR(20)* *VARCHAR(500)* *VARCHAR(200)* *INT* *INT* *VARCHAR(50)*	Ausência de chave primária. Existência de muitos *NULL* – Identificar o sentido da sua utilização.

O *dataset* disponibilizado é composto por seis tabelas relacionais, sendo que duas delas são responsáveis pela agregação das reclamações e as restantes quatro referem-se aos atributos que as sustentam. Após

o estudo das relações existentes no *dataset*, foi desenvolvido um modelo multidimensional inicial (Figura 9) resultante do armazenamento dos dados fornecidos no *MySQL Workbench*. Este modelo representa a estrutura dos dados fornecidos, sem que estes sofressem qualquer alteração.

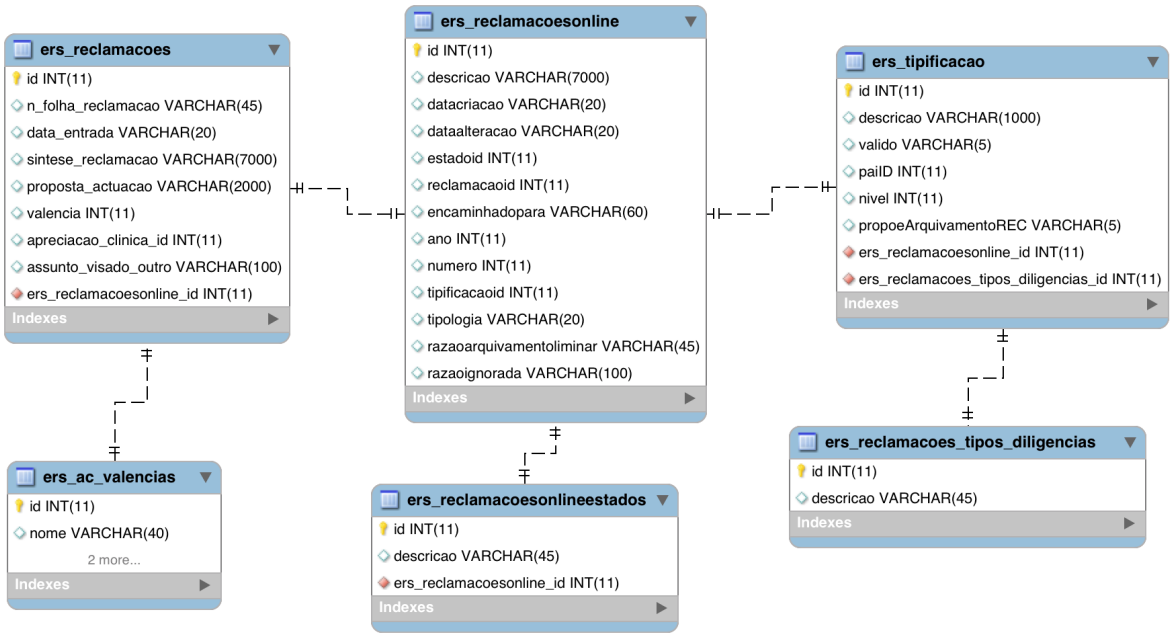

Figura 9 - Modelo multidimensional de dados não estruturados.

Através da Figura 9 é possível visualizar um modelo multidimensional resultante do *data warehouse* original. A tabela de factos, denominada "ers_reclamacoesonline" diz respeito a opiniões apresentadas via plataforma web. Esta tabela é composta por doze campos, sendo que para além de campos informacionais, apresenta ainda datas (data de criação e data de alteração), razão de arquivamento, razão de encaminhamento, tipologia, razão de ignoração e descrição da reclamação. Esta última é transcrita do portal tal e qual como o utilizador escreveu.

Como a própria Figura 9 demonstra, a tabela "ers_reclamacoesonline" é a tabela fundamental neste processo, sendo que todas as outras vão servir como complemento desta.

Relativamente às dimensões temos "ers_reclamacoes", "ers_ac_valencias", "ers_reclamacoesonlineestados", "ers_tipificacao" e "ers_reclamacoes_tipos_diligencias".

Sendo que a tabela "ers_reclamacoesonline" é a fundamental no processo segue-se em nível de importância a tabela "ers_reclamacoes". Nesta tabela podemos encontrar sete campos informativos relativos a reclamações provenientes de livros de reclamações. Toda esta informação já passou por processos de informatização, passando assim do formato papel para formato digital. É possível encontrar informação relativa ao número de folha do livro em que a reclamação está escrita, data de redação,

síntese da reclamação e uma pequena apreciação por parte do responsável analisador da reclamação. Nas reclamações em papel o processo de tratamento da informação difere um pouco das reclamações online, pois a reclamação não é totalmente transcrita do papel, passando assim por um processo de triagem em que o responsável por este processo revê a reclamação e faz uma síntese da mesma, deixando uma opinião de atenuação.

A tabela "ers_ac_valencias" é composta por um campo informacional que diz respeito à descrição da valência, isto é, ao departamento médico onde a reclamação pode ser inserida.

A tabela "ers_reclamacoesonlineestados" é igualmente composta por um campo informacional que diz respeito ao estado atribuído a cada reclamação.

A tabela "ers_tipificacao" é composta por cinco campos informacionais, contendo informação relativa à categorização que é dada à reclamação em causa, ou seja, a cada reclamação é atribuído um tipo com base na análise feita da sua descrição. A este tipo de reclamação acresce um campo referente à proposta de arquivamento caso a reclamação não cumpra os requisitos de aceitação.

Por último a tabela "ers_reclamacoes_tipos_diligencias" é composta por um campo informacional referente à importância dada à reclamação conforme o seu tipo. Esta importância pode, por exemplo, corresponder à instauração de um processo disciplinar ao órgão em questão.

4.2 Classificação do problema

O problema central deste projeto de investigação está diretamente relacionado com o processo de gestão de reclamações em centros prestadores de cuidados de saúde. Como foi anteriormente mencionado a ERS é responsável pela gestão e análise das reclamações, recebendo diariamente um grande volume de dados, muitos deles mal estruturados sem análise possível. Por outro lado, grande parte deste volume diz respeito a problemas reais, bem como críticas construtivas e elogios ao serviço disponibilizado pelo médico e do centro prestador de cuidados de saúde.

Desta forma, o problema foca-se na qualidade da reclamação efetuada pelo utente, pretendendo-se uma melhoria dos padrões de resposta bem como uma análise de conhecimento ao nível dos Sistemas de Informação (SI) aplicados à saúde.

Numa perspetiva de melhoria contínua, é fundamental para ambas as partes, uma boa qualidade dos dados de forma a facilitar a identificação do problema reportado na reclamação. Esta fácil identificação pode ter dois extremos, o extremo positivo, no caso de estarmos presentes a consecutivos elogios a um determinado médico pelo serviço prestado em que o médico pode ser sujeito a uma recompensa pelo seu excelente trabalho. Por outro lado, caso estejamos presentes a consecutivas denúncias a um

determinado médico, estamos presentes a um extremo negativo em que o médico pode ser sujeito a um processo de investigação com aplicação de sanções caso o problema seja identificado e provado.
Com base nestes pontos identificados, podemos concluir que este tema é de elevada importância, sendo que a qualidade da reclamação pode ter bastante impacto na sua análise e resolução.

4.3 Desenho e solução

4.3.1 Seleção e agregação de dados de investigação

Após a análise efetuada ao volume de dados fornecidos, foram estabelecidas algumas alterações, mantendo sempre o foco do projeto bem como a concordância com os orientadores. A alteração fundamental neste projeto de dissertação passou pela transformação de um modelo com uma tabela de factos principal (ers_reclamacoesonline) para um modelo com duas tabelas de factos distintas (Reclamacao_papel e Reclamacao_online). Desta forma, é possível uma melhor distinção das reclamações, percebendo aquelas que efetivamente têm mais qualidade e são mais percetíveis. Uma outra alteração, não menos importante que a anterior, corresponde ao facto de terem sido descartados atributos desnecessários neste tema, como por exemplo atributos de texto, sendo o mais notório o atributo referente à reclamação transcrita do livro ou agregada do próprio portal de submissão. Com base no grande volume de dados, o facto de não serem contemplados estes atributos é possível obter uma melhor performance no que toca ao processamento da base de dados. Juntamente com as alterações contempladas, foi agregada uma nova tabela para o tempo com base nas datas referenciadas na submissão das reclamações. Por último, para uma melhor consistência dos dados, apenas foram contemplados dados que obtinham correspondência entre tabelas.
Assim, estavam reunidas todas as condições necessárias para o início do tratamento dos dados estabelecidos como essenciais para o projeto de investigação.

5 DESENVOLVIMENTO DA SOLUÇÃO

Este capítulo reflete a parte mais prática do projeto de investigação, estando dividido em cinco secções, entre elas a preparação dos dados, o processo de *Extract, Transform e Load* (ETL), a criação do cubo *Online Analytical Processing* (OLAP), a introdução ao *Power BI* e a elaboração de modelos de regressão e classificação. Na primeira secção é possível identificar todas as etapas de preparação dos dados. Na secção seguinte é apresentado todo o processo ETL bem como o tratamento a que os dados foram sujeitos. Na secção subsequente é possível observar o processo de criação do cubo OLAP, responsável pela disponibilização dos dados para análise. Na secção seguinte é apresentado o processo de criação de dashboards bem como a sua análise. Por último, é apresentado uma proposta de desenvolvimento de *Data Mining* através de regressão e classificação.

5.1 Preparação dos dados

A preparação dos dados em estudo é fundamental para a obtenção de sucesso no desenvolvimento do projeto, sendo importante precaver as lacunas que possam existir. Desta forma, com base no que fora explicado anteriormente no processo de seleção e agregação dos dados é importante reforçar a ideia que estes dados se apresentavam como dados não estruturados, com demasiados erros e com alguma falta de informação.

Para uma melhor perceção do tratamento efetuado, foi desenvolvido um modelo multidimensional, conforme a Figura 10, onde algumas das alterações anteriormente referidas estão presentes.

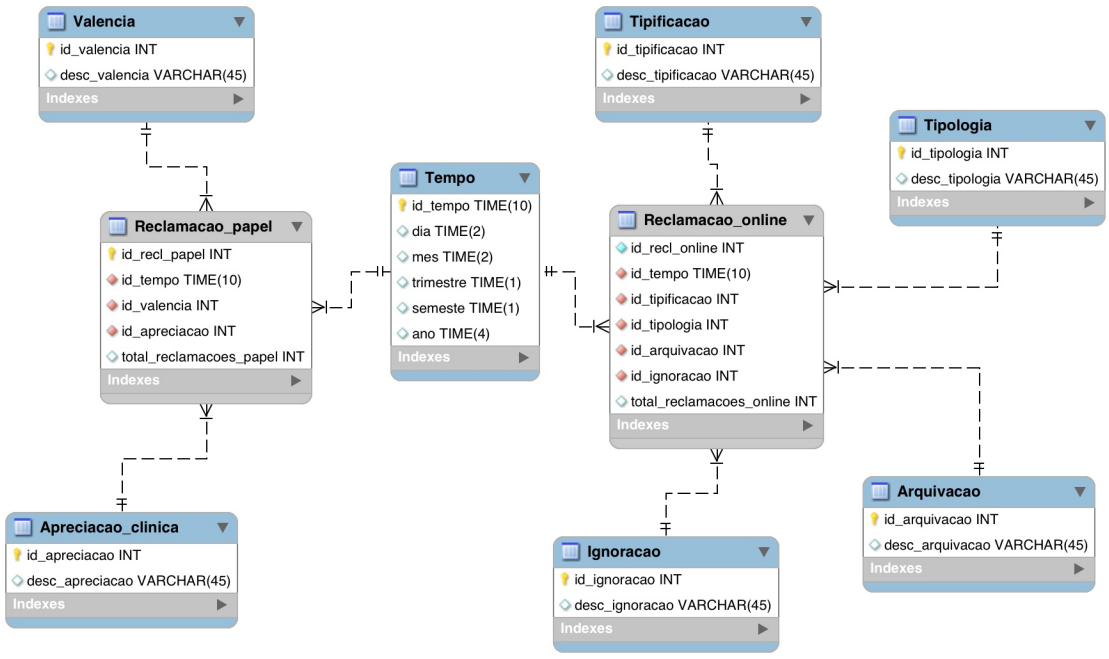

Figura 10 - Modelo multidimensional de dados estruturados.

Como se pode observar na Figura 10, as tabelas de factos "Reclamacao_papel" e "Reclamacao_online" são constituídas apenas por *id's* provenientes das dimensões que as sustentam, sendo elas "Valencia" e "Apreciacao_clinica" para as reclamações em papel e "Tipificacao", "Tipologia", "Estado" e "Ignoracao" para as reclamações online. Existe ainda uma dimensão comum às duas tabelas de factos, a dimensão "Tempo", dotada de diferentes tipos de granularidade.

As tabelas relacionadas diretamente com as tabelas de factos "Reclamacao_papel" e "Reclamacao_online" são compostas por *id's* e as respetivas descrições pelas quais existe relação com as reclamações. Todas as outras descrições que não vão de encontro com estes parâmetros não são contempladas.

Após o desenvolvimento do modelo multidimensional foi identificada a necessidade de criação de uma base de dados auxiliar de modo a que fosse possível estabelecer a associação dos dados presentes nas tabelas de dados não estruturados mencionadas anteriormente "ers_reclamacoes" e "ers_reclamacoesonline" com as tabelas necessárias para satisfazer o modelo criado.

Primeiramente foram identificados os dados de correspondência direta, ou seja, dados que já se encontravam em tabelas separadas. O resultado desta primeira identificação são as tabelas "Valencia", "Estado" e "Tipificacao". Através de uma *query* foi então possível a inserção dos dados pretendidos (id e descricao) numa base de dados criada no *SQL Server* previamente preparada com a respetiva estruturação de dados. É de realçar, uma vez mais, que apenas os dados que se intersetam foram considerados neste estudo.

A Figura 11 demonstra algumas das valências atribuídas às reclamações, num total de 43 registos, identificando a área clínica em que cada reclamação incide.

id_valencia	desc_valencia
3	Radiologia
4	Medicina física e de reabilitação
5	Análises clínicas / patologia clínica
6	Anatomia patológica
7	Electroencefalografia
8	Endoscopia gastrenterológica
9	Especialidades médico cirúrgica
10	Análises clínicas / por farmacêuticos
11	Cardiologia
12	Medicina nuclear
13	Neurofisiologia
14	Otorrinolaringologia
15	Pneumologia e imunoalergologia
16	Urologia
17	Cirurgia / prestação de cuidados
18	Diálise
19	SIGIC
20	Dentistas
21	Psicologia Clínica
22	Pediatria
23	Ginecologia
24	Ortopedia

Figura 11 - Listagem de valências.

A Figura 12, com 6 registos, demonstra os estados possíveis utilizados na classificação de cada etapa do processamento das reclamações.

id_estado	desc_estado
1	Anexada a Processo
2	Gerou Rec
3	Arquivada Liminarmente
4	Inserida
5	Reencaminhada
6	Ignorada

Figura 12 - Listagem de estados.

A Figura 13 demonstra algumas caracterizações atribuídas a cada reclamação, sendo que esta especificação do tema da reclamação é importante para a distinção de área incidente. Numa primeira análise, contando com registos idênticos, existem 30 registos consideráveis.

id_tipificacao	desc_tipificacao
1	Acesso
2	Acesso a cuidados primários do SNS
4	Não atribuição de médico de família
5	Inexistência de enfermeiro assistente
6	Não registo imediato do utente no sistema de informação do pedido de consulta
8	Não registo imediato do utente no sistema de informação do pedido de tratamento
9	Impossibilidade de livre escolha da unidade de cuidados primários
11	Acesso a cuidados hospitalares do SNS
12	Incumprimento de TMRG
13	Incumprimento das regras do SIGIC
16	Não registo imediato do utente no sistema de informação do pedido de exame médico
18	Impossibilidade de livre escolha do médico assistente
19	Impossibilidade de livre escolha do enfermeiro assistente
20	Outros Assuntos
21	Acesso a cuidados continuados
22	Incumprimento das regras de referenciação
23	Impossibilidade de livre escolha da unidade de cuidados continuados
24	Acesso a cuidados convencionados SNS
25	Impossibilidade de livre escolha de entidade convencionada
26	Cobrança de taxas moderadoras além das legalmente estabelecidas
27	Incumprimento dos TMRG definidos legalmente para cada prestação concreta
28	Entraves causados pela entidade convencionada

Figura 13 - Listagem de tipificações.

Para os dados da tabela "Apreciacao_clinica" foi desenvolvida uma *query* que verificasse todos os valores possíveis existentes no atributo "apreciacao_clinica_id" da tabela "ers_reclamacoes". No carregamento dos dados para a nova base de dados, foi atribuída aos valores uma descrição. As descrições definidas são "Considerado" para valores iguais a 0 e "Não Considerado" para valores iguais a 1.

A Figura 14 demonstra as apreciações disponíveis para exploração, indicando desta forma se a reclamação em causa é passível de análise ou não. Através deste atributo é possível ao responsável pela primeira análise das reclamações em papel indicar se existe relevância na mesma e se efetivamente é indicada para estudo.

id_apreciacao	desc_apreciacao
0	Não considerado
1	Considerado

Figura 14 - Listagem de apreciações clínicas.

Relativamente aos dados da tabela "Tipologia" foi aplicada uma *query* que verificasse todas as descrições possíveis e distintas existentes no atributo "tipologia" da tabela "ers_reclamacoesonline". Desta forma, aquando do carregamento dos dados foi possível a atribuição de um id a cada descrição encontrada.

A Figura 15, com 6 registos, demonstra as tipologias atribuídas às reclamações onde é possível perceber que não é tudo classificado como reclamação, mas também como denúncia, elogio, exposição e sugestões. A classificação como "outras" refere-se a tipologias que não se enquadram nas restantes. É possível assim identificar a má qualidade dos dados quando existem tipologias sem significado no contexto desta classificação.

id_tipologia	desc_tipologia
1	Denúncia
2	Elogio/Louvor
3	Exposição
4	Outras
5	Reclamação
6	Sugestão

Figura 15 - Listagem de tipologias.

Por último, relativamente à tabela "Ignoracao" foi aplicada uma *query* que verificasse todas as descrições possíveis e distintas existentes no atributo "razaoaignoracao" da tabela "ers_reclamacoesonline". Desta forma, aquando do carregamento dos dados foi possível a atribuição de um id a cada descrição encontrada conforme a Figura *16*.

id_ignoracao	desc_ignoracao
1	Anexada manualmente ao processo REC_4970/2014
2	Anónima
3	Considerou-se extemporânea por redundante a anexação ao processo REC_592/2014
4	Denúncia Anónima; inserido pedido numericerno de fiscalização.
5	Desconhecida
6	Encaminhada numericernamente - DAR e GGQ
7	Encaminhada numericernamente - DAU e GGQ
8	Encaminhada para o DAR
9	Encaminhada por e-mail para o DAR
10	Encaminhado para MP
11	Encaminhado para o DAR
12	Encaminhado para o DQF
13	Estupidez natural
14	EXP para o DAR
15	Foi aberta REC manualmente REC 1072/2014
16	Fora de âmbito
17	Repetição
18	Teste
19	Transformada manualmente em REC_6009/2014
20	Transformada manualmente: REC_6010/2014
21	Transformada manualmente: REC_6060/2014

Figura 16 - Listagem de razões de ignoração.

5.2 Processo ETL

O processo ETL é o processo responsável pela extração de dados de várias fontes, limpeza, otimização e inserção desses mesmos dados numa base de dados de destino, também designada por DW (*data warehouse*) (Negash, 2004).

Nesta fase do projeto através do módulo *SQL Server Integration Services* (SSIS) do *visual studio data tools* foi desenvolvido o modelo apresentado na Figura 17, organizando assim todo o processo de transformação e carregamento dos dados.

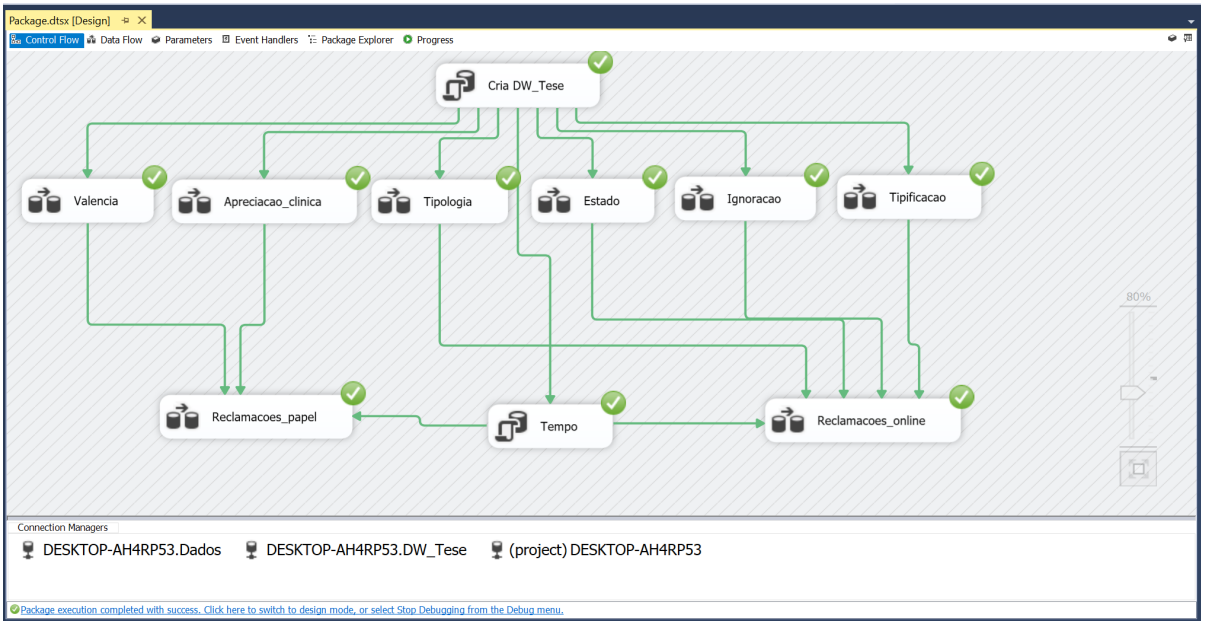

Figura 17 - Modelo desenvolvido no visual studio data tools para o processo ETL.

Este processo pode ser dividido em cinco fases fundamentais, sendo elas a criação da base de dados, o tratamento dos dados, o carregamento dos dados, o tratamento dos factos e o carregamento dos factos.

Numa primeira fase, foi desenvolvida uma *query* (Cria DW_Tese) responsável pela criação de todas as tabelas necessárias para o estudo. Esta *query* resultou da análise prévia dos dados, garantindo que no carregamento dos dados não existissem problemas de incoerência, ou seja, que existisse concordância de tipos, nomeadamente tamanho e precisão do campo bem como o tipo correspondente (numérico ou texto). A tabela tempo, também desenvolvida através de uma *query* é responsável pelo carregamento de todas as datas necessárias para o estudo, tendo por base o intervalo de tempo considerado no *dataset* fornecido.

Com esta fase concluída, foi então necessário o tratamento dos dados previamente carregados e organizados numa base de dados auxiliar. Nesta fase, os dados não sujeitos a tratamento foram os dados

respetivos às valências e aos estados. Todos os outros sofreram alterações necessárias para este desenvolvimento, seguidamente apresentadas por tópicos para melhor compreensão:

- Apreciacao_clinica
    - Colocação da descrição "Não considerado" para dados com id 0;
    - Colocação da descrição "Considerado" para dados com id 1.
- Tipologia
    - Colocação da descrição "Outras" para dados com a descrição "*Null*" e "-1".
- Ignoracao
    - Colocação da descrição "Desconhecida" para dados com a descrição "*Null*" e " ";
    - Colocação da descrição "Fora de âmbito" para dados com a descrição "Sem fundamento";
    - Colocação da descrição "Anónima" para dados com a descrição "Anónimo", "Reclamação Anónima", "Anónima, sem matéria para analisar" e "Denúncia Anónima; inserido pedido interno de fiscalização";
    - Colocação da descrição "Repetição" para dados com a descrição "repetição".
- Tipificacao
    - Colocação da descrição "Outros Assuntos" para dados com a descrição "Outros", "Outro", "Outros temas" e "*Null*".

Nesta fase, os dados já tratados foram então carregados para a base de dados final sustentando com factos as seguintes tabelas: "Reclamacoes_papel" e "Reclamacoes_online".
Para este processo de carregamento de factos foi necessário o cruzamento dos dados do *dataset* inicial com os dados sujeitos ao tratamento apresentado anteriormente.
Para o carregamento dos factos das "Reclamacoes_papel" foram cruzados os dados primordiais com as tabelas "Valencia", "Apreciacao_clinica" e "Tempo". Para além do tratamento efetuado anteriormente que teve de ser aplicado novamente nesta etapa, foi efetuado o seguinte tratamento:

- Valencia
    - Não existindo correspondência para valências com id inferior a 3 e superior a 44, foi atribuído a estes dados o id 80, referente à descrição "Desconhecido".
- Tempo
    - Tratamento da data com base na hierarquia já apresentada, dia, mês e ano. O id da data é representado pela própria data.

Para o carregamento dos factos das "Reclamacoes_online" foram cruzados os dados primordiais com as tabelas "Tipologia", "Estado", "Ignoracao", "Tipificacao" e "Tempo". Para além do tratamento efetuado anteriormente que teve de ser aplicado novamente nesta etapa, foi efetuado o seguinte tratamento:

- Tipificacao
    - Não existindo correspondência para tipificações com id "*Null*" e "-1", foi atribuído a estes dados o id 296, referente à descrição "Outros Assuntos".
- Tempo
    - Tratamento da data com base na hierarquia já apresentada, dia, mês e ano. O id da data é representado pela própria data.

Após o tratamento dos dados, salvaguardando todas as hipóteses de erro, foi possível efetuar o processamento deste modelo inserindo 12255 registos para as reclamações em papel (Figura 18) e 1799 registos para as reclamações online (Figura 19).

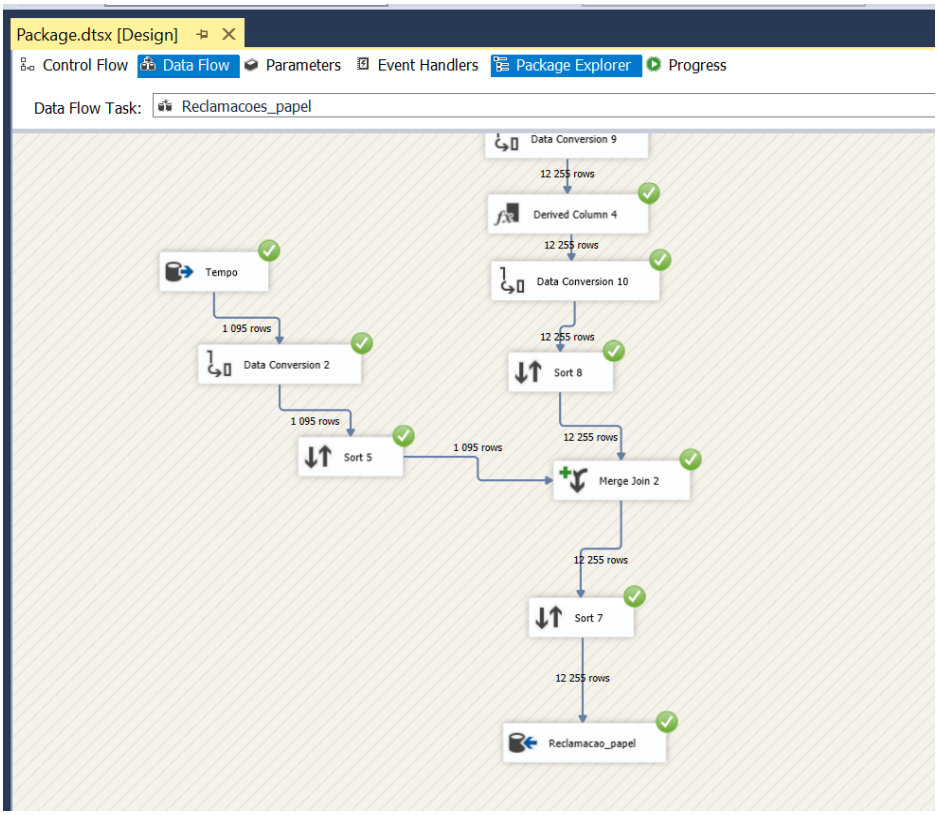

Figura 18 - Total de registos inseridos para as reclamações em papel.

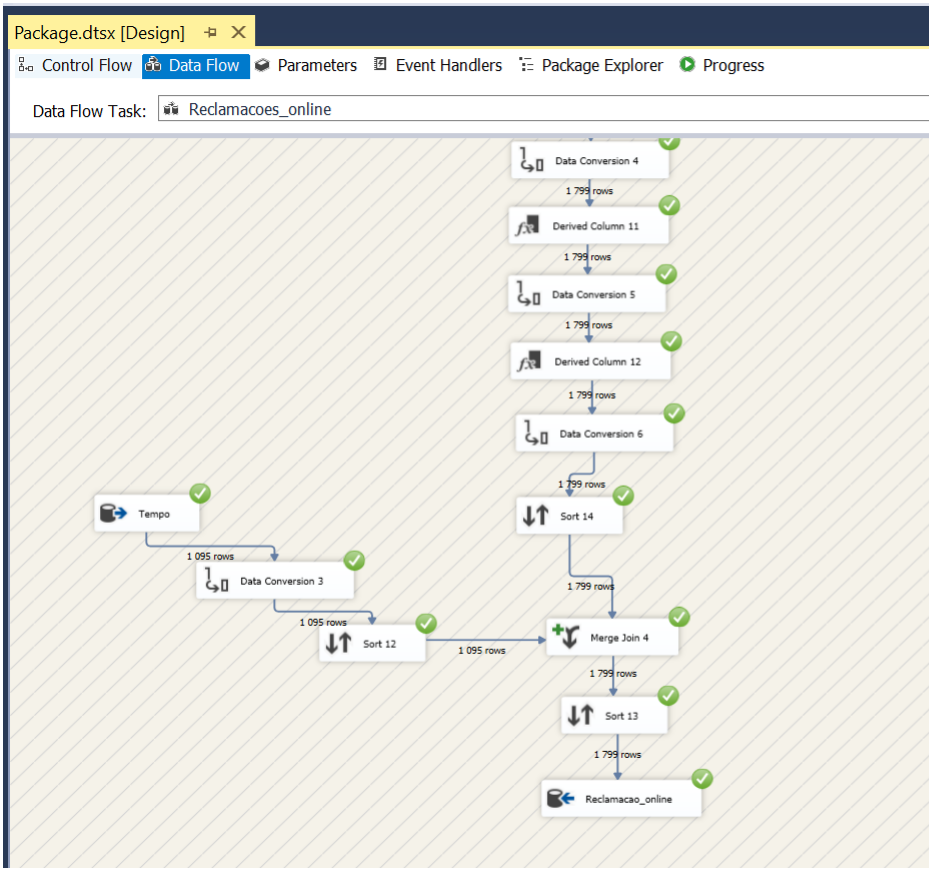

Figura 19 - Total de registos inseridos para as reclamações online.

Com o processo de tratamento e carregamento dos dados desenvolvido, com um total de 14054 registos de reclamações, estão reunidas condições para o desenvolvimento do cubo no módulo de *SQL Server Analysis Services* (SSAS) do *visual studio data tools*, demonstrado na Secção 5.3.

5.3 Criação do cubo

O cubo OLAP, considerado por Berson & Smith (Berson & Smith, 1997) uma estrutura de dados capaz de fornecer uma análise rápida dos dados através de vistas multidimensionais, permite a identificação de padrões e tendências nos dados. Estes sistemas são muito usuais em modelos de dados multidimensionais como o utilizado no projeto de investigação, *data warehouse*.

Nesta etapa foi desenvolvido o cubo OLAP, que será posteriormente considerado na exploração dos dados através de *dashboards* na ferramenta *Power BI*.

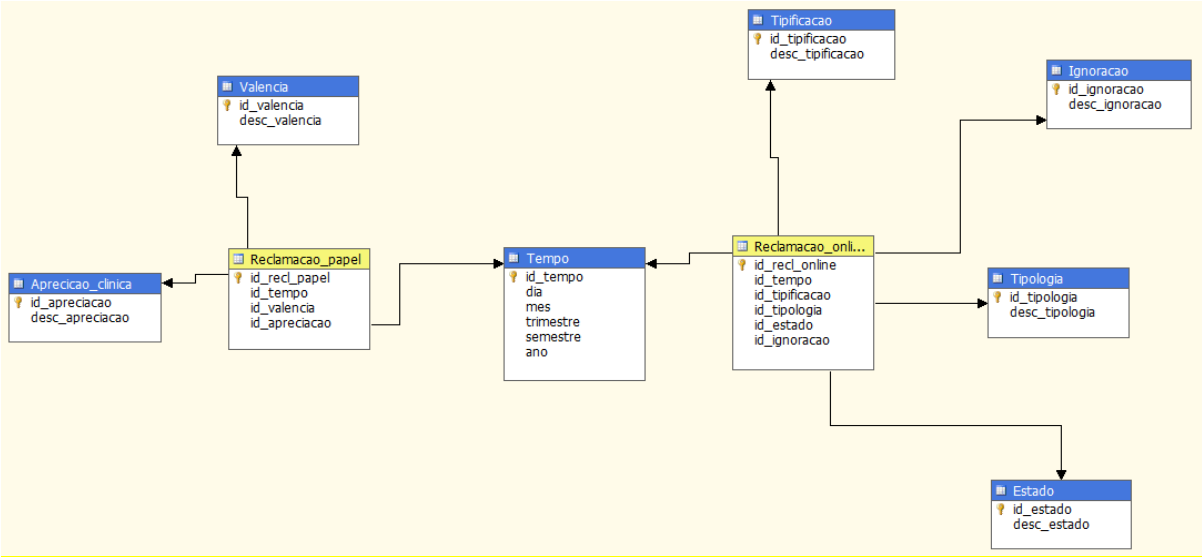

Figura 20 - Modelo multidimensional do cubo OLAP.

No desenvolvimento do cubo foi então criada uma ligação ao *data warehouse* anteriormente criado, como é possível visualizar na Figura 20 o modelo multidimensional representante é exatamente igual ao modelo multidimensional considerado nas etapas anteriores. Com as ligações estabelecidas foi necessário desenvolver as hierarquias necessárias, sendo que a mais importante neste processo é a hierarquia da dimensão tempo, pois todas as outras dimensões carecem apenas da descrição do dado em causa. Neste sentido, a hierarquia considerada para a dimensão tempo é:

- Ano;
- Semestre;
- Trimestre;
- Mês;
- Dia.

Com as hierarquias de todas as dimensões criadas foi tempo de processar o cubo até então desenvolvido, sendo que o resultado do seu processamento é possível visualizar na Figura 21.

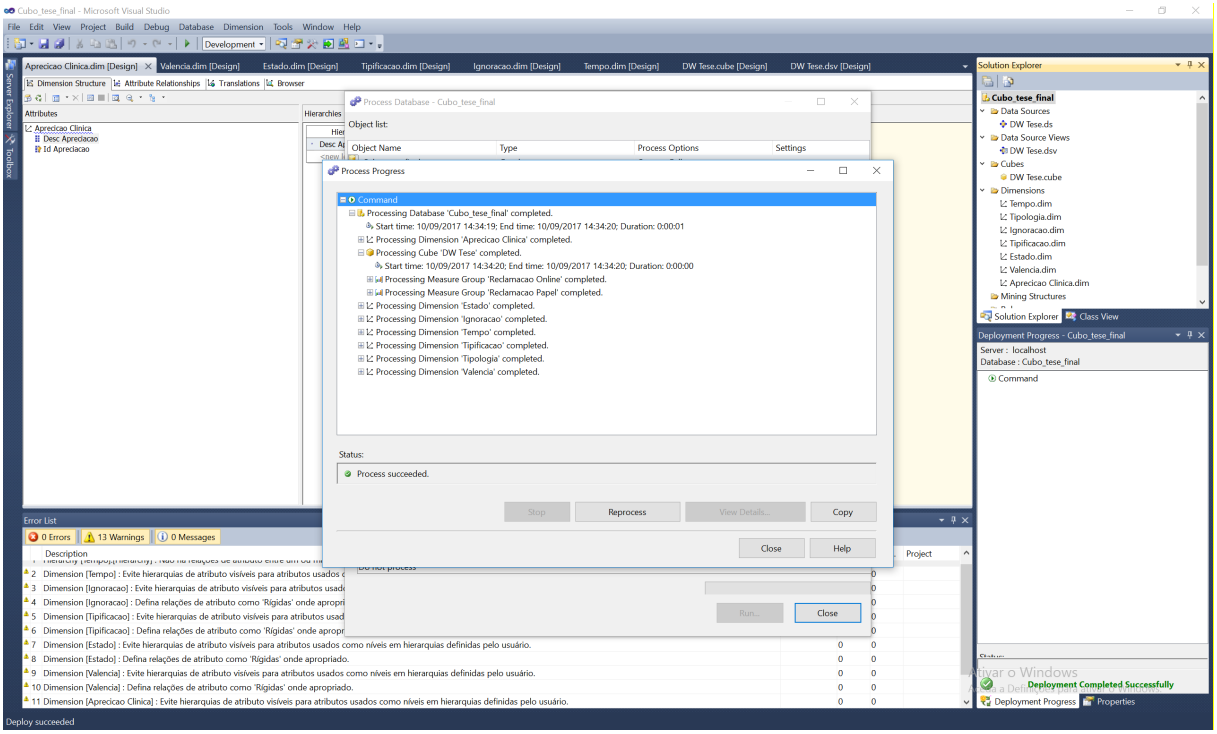

Figura 21 - Resultado do processamento do cubo OLAP.

5.4 Introdução ao Power BI

O Power BI desenvolvido pela Microsoft é considerado como um conjunto de ferramentas de análise de negócios capaz de oferecer ao utilizador uma visão global em tempo real de todo o negócio em estudo num único local com base nas métricas definidas. (Microsoft, 2010)

Existem inúmeras ferramentas dotadas desta capacidade de análise do negócio em tempo real. No entanto, no contexto deste projeto de investigação e com base em outros projetos já efetuados foi escolhida a ferramenta *Power Bi* para o desenvolvimento de *dashboards* na etapa que se segue.

Um dos fatores mais importante no processo de escolha da ferramenta a utilizar foi o contacto já obtido com a ferramenta, mas também foi tido em consideração a facilidade de disponibilização de informação relevante para o processo de tomada de decisão, apresentando-se como uma ferramenta bastante intuitiva em toda a sua utilização.

5.4.1 Processo de criação de dashboards

Nesta fase de processamento de *dashboards* foi preponderante o desenvolvimento de *dashboards* gerais de forma a interagir com os dados, percebendo assim qual a utilidade e usabilidade dos dados em estudo. Neste sentido foram desenvolvidos *dashboards* para os dois tipos de reclamações como é possível visualizar na Figura 22, Figura 23 e Figura 24.

Na Figura 22, referente à análise das reclamações em papel, é possível identificar os filtros temporais desenvolvidos que demonstram os anos em que estes dados incidem, sendo eles 2013, 2014 e 2015. Sem a aplicação de filtros, podemos ainda verificar na Figura 22 que existem 12257 registos referentes às reclamações em papel, sendo que a valência mais representativa com 70,96% de incidência é a valência "Desconhecido". As valências subsequentes com percentagens muito inferiores, mas não menos importantes a esta são: "Ortopedia" com 4,76%, "Medicina numericerna" com 2,68% e "Oftalmologia" com 2,55%. Por último é possível identificar que todos os registos têm uma apreciação clínica de "Considerado".

De uma forma geral, podemos considerar que o grave problema nas reclamações em papel é a identificação das valências, nomeadamente nos registos em que esta é desconhecida, tornando-se muito difícil uma avaliação exata da área com maior número de registos incidentes para rápida atuação.

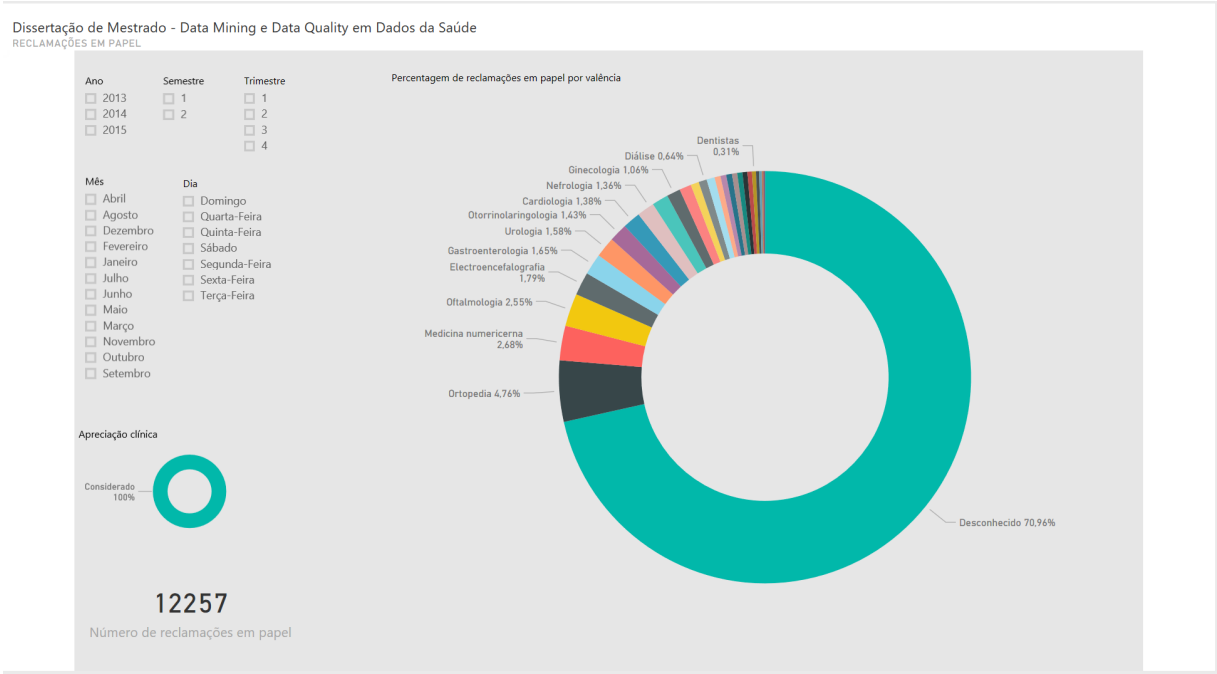

Figura 22 - Dashboard geral referente às reclamações em papel.

Nas imagens seguintes, Figura 23 e Figura 24, que dizem respeito à análise das reclamações online podemos verificar os filtros temporais, onde é visível que ao contrário das reclamações em papel apenas constam dados referentes aos anos 2014 e 2015. Sem a aplicação de filtros, podemos ainda verificar na Figura 23 e na Figura 24 que existem 1799 registos referentes às reclamações online.

Na Figura 23, podemos visualizar que a tipificação mais representativa é "Outros assuntos" com 56,09% das incidências. As tipificações subsequentes com percentagens muito inferiores, mas não menos importantes são: "Qualidade dos Cuidados de Saúde" com 7,95%, "Direitos dos Utentes" com 6%, "Tempos de espera" com 5,95% e "Livro de reclamações" com 3,67%.

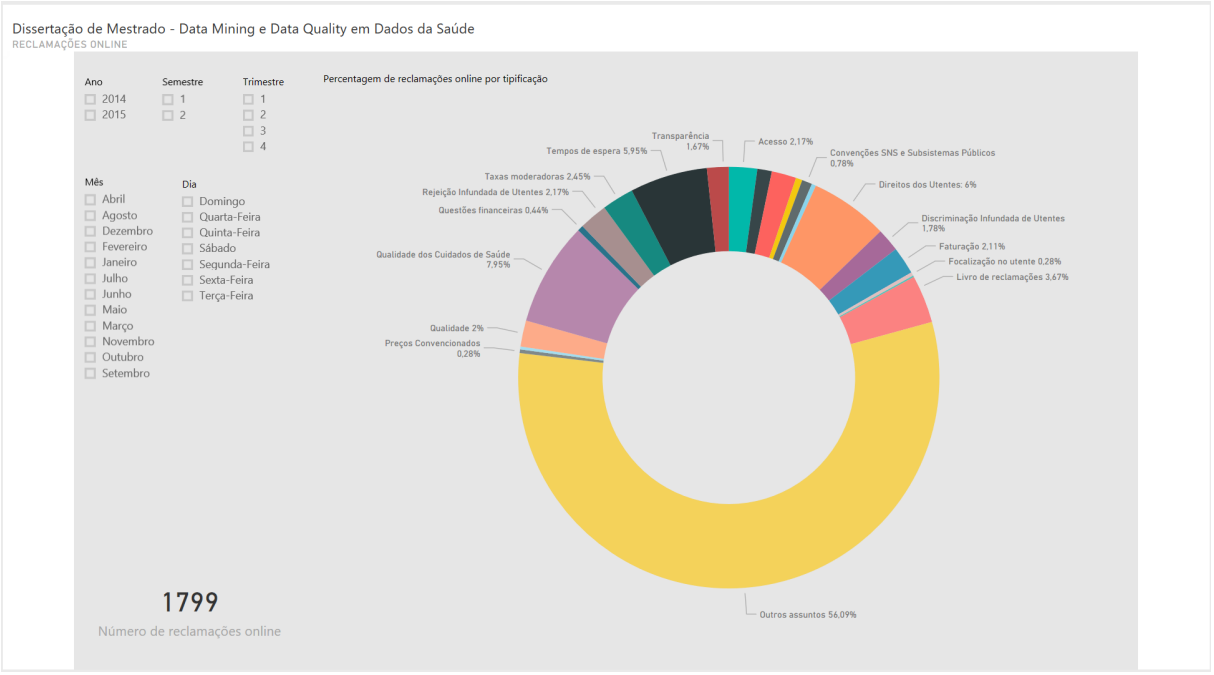

Figura 23 - Dashboard geral referente às reclamações online (parte 1).

Na Figura 24, podemos visualizar que os estados mais representativos são: "Gerou Rec" com 44,02% e "Inserida" com 32.96% das incidências. Os estados subsequentes com percentagens inferiores, mas não menos importantes são: "Ignorada" com 11.62%, "Arquivada Liminarmente" com 6% e "Anexada a Processo" com 4,95%. Para além dos estados, podemos visualizar ainda os motivos de ignoração, sendo o mais representativo a "Desconhecida" com 88,38% das incidências. Um outro motivo, não menos importante é a "Repetição" com 9,95%. Por último, relativamente à tipologia da reclamação podemos identificar "Outras" como a mais incidente com 844 registos, seguida da "Reclamação" com 745, da "Denúncia" com 122 e da "Exposição" com 83 registos.

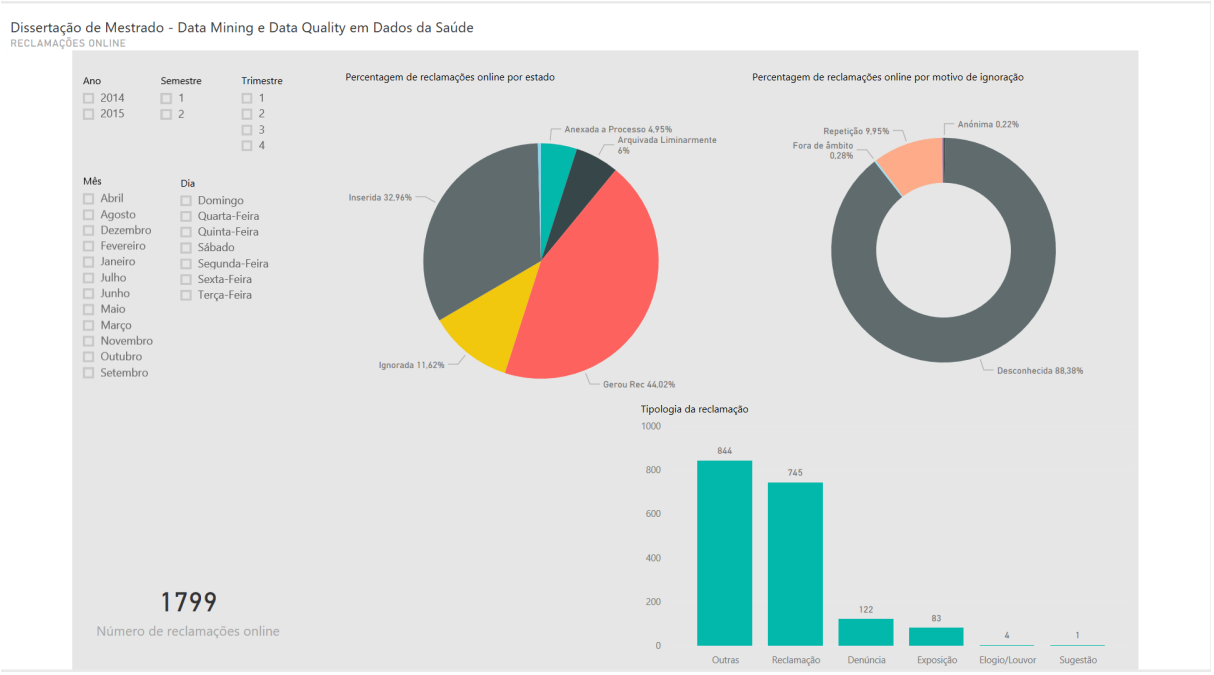

Figura 24 - Dashboard geral referente às reclamações online (parte 2).

Relativamente às reclamações online, com base nos *dashboards* apresentados anteriormente podemos referir que prevalecem as designações "Desconhecida" e "Outros assuntos", tornando os dados pouco identificativos e relevantes para uma atuação eficiente de medidas corretivas.

5.4.2 Análise dos dashboards desenvolvidos

Nesta fase foram explorados os *dashboards* apresentados na secção 5.4.1, evidenciando de certa forma uma má qualidade dos dados que os suportam. Neste sentido, com a aplicação dos filtros é possível obter uma perceção temporal mais pormenorizada, como por exemplo a altura do ano em que existe mais afluência de reclamações registadas.

Este estudo centrou-se essencialmente na análise do ano, semestre, trimestre, mês e dia com mais reclamações registadas. De seguida, podemos encontrar 5 análises, das quais 3 são relativas às reclamações em papel e 2 relativas às reclamações online.

Reclamações em papel

A primeira análise desenvolvida referente ao ano 2013 em que apenas existem registos para o 2º Semestre, conta com um total de 9 reclamações sendo que 7 dizem respeito ao período selecionado, como pode ser visualizado na Figura 25. Neste período prevalece a valência "Desconhecido" com 28,57% dos registos que diz respeito a 2 registos, as restantes valências apresentam 1 registo cada.

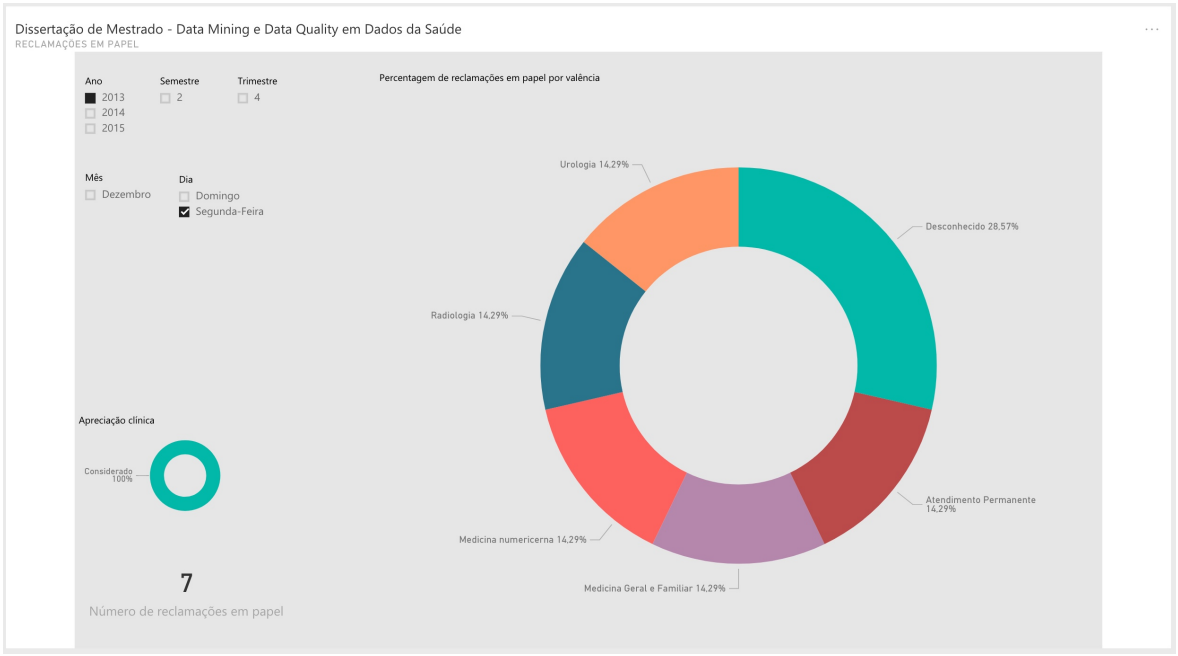

Figura 25 - Análise das reclamações em papel em 2013.

A segunda análise desenvolvida referente ao ano 2014 em que existem registos do ano completo, conta com um total de 11013 reclamações sendo que 322 dizem respeito ao período selecionado, como pode ser visualizado na Figura 26. Neste período prevalece a valência "Desconhecido" com 72,67% dos registos que diz respeito a 234 registos. A subsequente valência, não menos importante é "Ortopedia" com 4,04% que diz respeito a 13 registos.

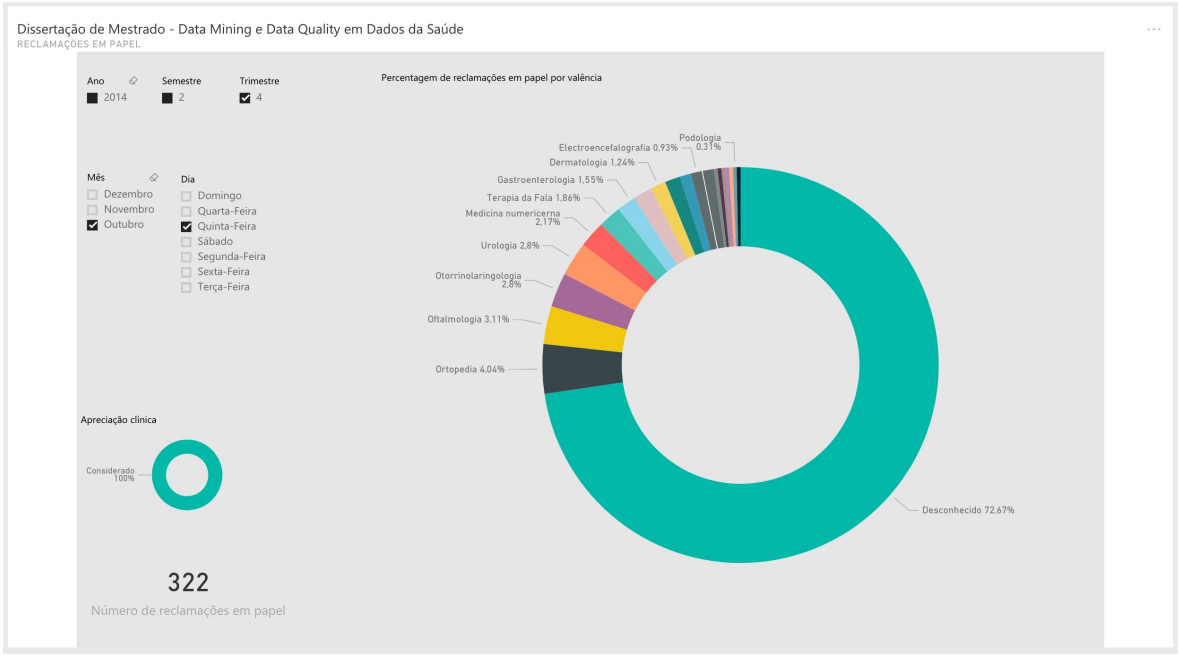

Figura 26 - Análise das reclamações em papel em 2014.

A terceira análise desenvolvida referente ao ano 2015 em que apenas existem registos para o 1º Semestre, conta com um total de 1235 reclamações sendo que 282 dizem respeito ao período

selecionado, como pode ser visualizado na Figura 27. Neste período prevalece a valência "Desconhecido" com 70,57% dos registos que diz respeito a 199 registos. A subsequente valência, não menos importante é "Ortopedia" com 3,55% que diz respeito a 10 registos.

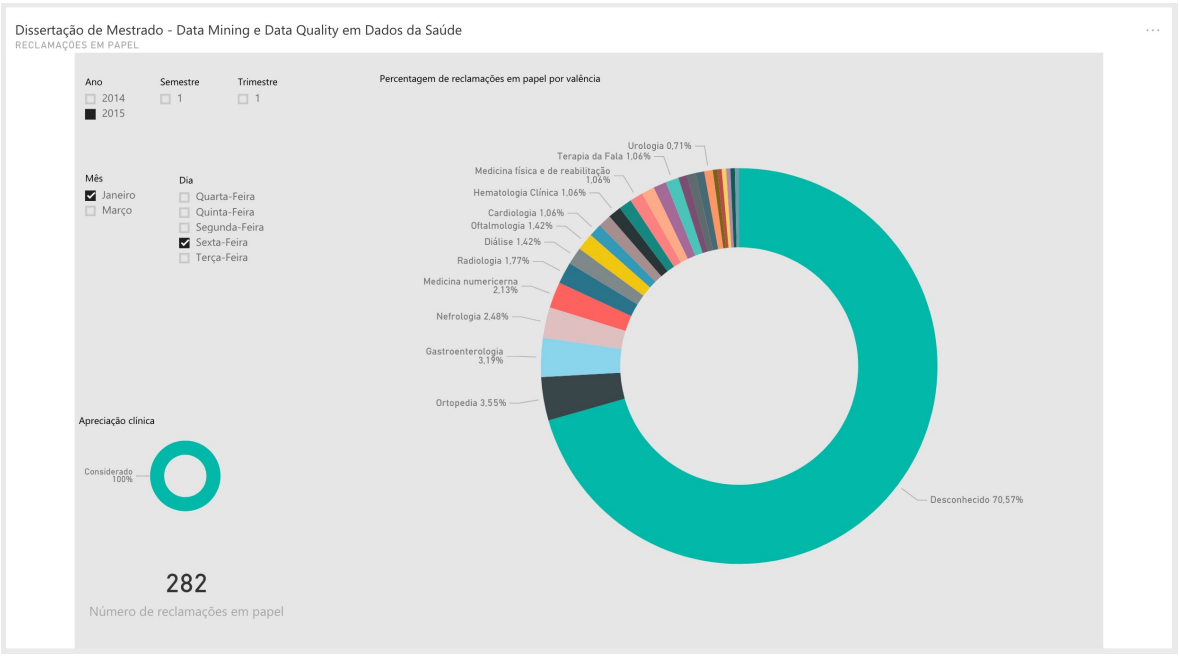

Figura 27 - Análise das reclamações em papel em 2015.

Relativamente às análises das reclamações em papel apresentadas anteriormente, com base na estação do ano em causa (Outono/Inverno) podemos considerar que a afluência de reclamações poderá dever-se ao início do frio e consequência disso o aparecimento das primeiras gripes. Segundo a DGS a gripe é uma doença sazonal que se manifesta principalmente durante o Inverno (Freitas, 2015). Contudo, são conclusões abstratas, pois a grande afluência dos registos apresenta valência desconhecida. Nestes casos, deveria ser considerada a possibilidade de alargamento de valências dando mais especificidade a este tema, reduzindo assim o elevado número de registos pouco conclusivos.

Reclamações online

A primeira análise desenvolvida referente ao ano 2014 em que existem registos do ano completo, conta com um total de 1511 reclamações sendo que 42 dizem respeito ao período selecionado, como pode ser visualizado na Figura 28 e na Figura 29. Neste período prevalece:

- Tipificação
    - "Outros assuntos" com 19,05% dos registos (8 registos);
    - "Tempos de espera" com 14,29% dos registos (6 registos).

- Estado
    - "Inserida" com 90,48% dos registos (38 registos).
- Ignoração
    - "Desconhecida" com 95,24% dos registos (40 registos)
- Tipologia
    - "Reclamação" com 29 registos.

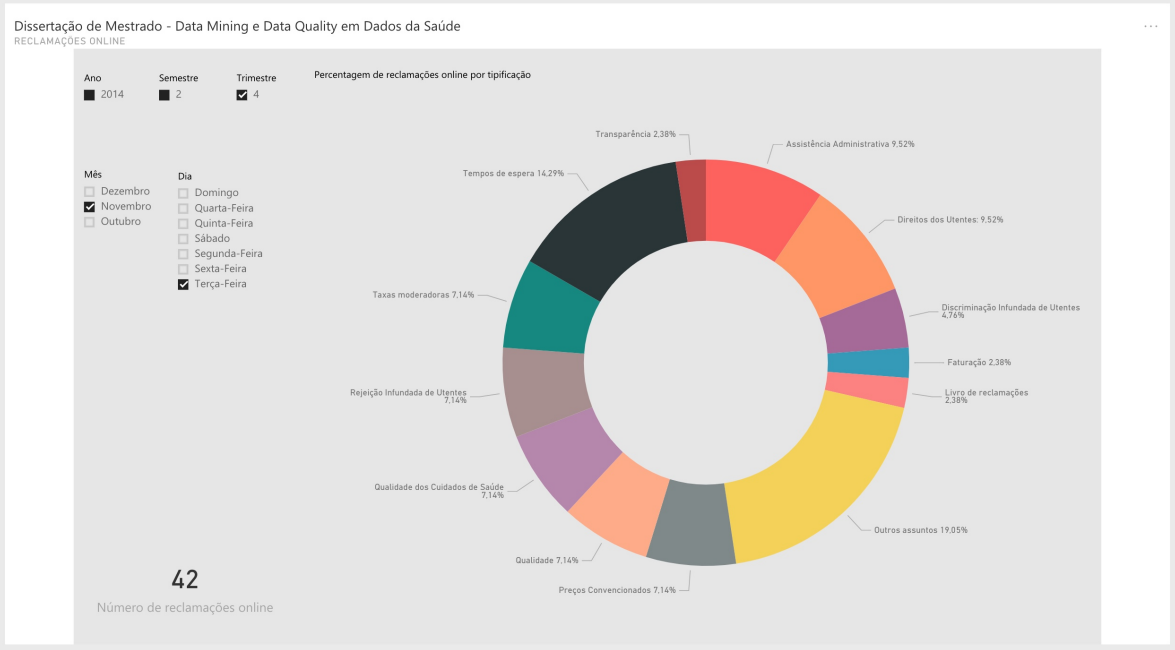

Figura 28 - Análise das reclamações online em 2014.

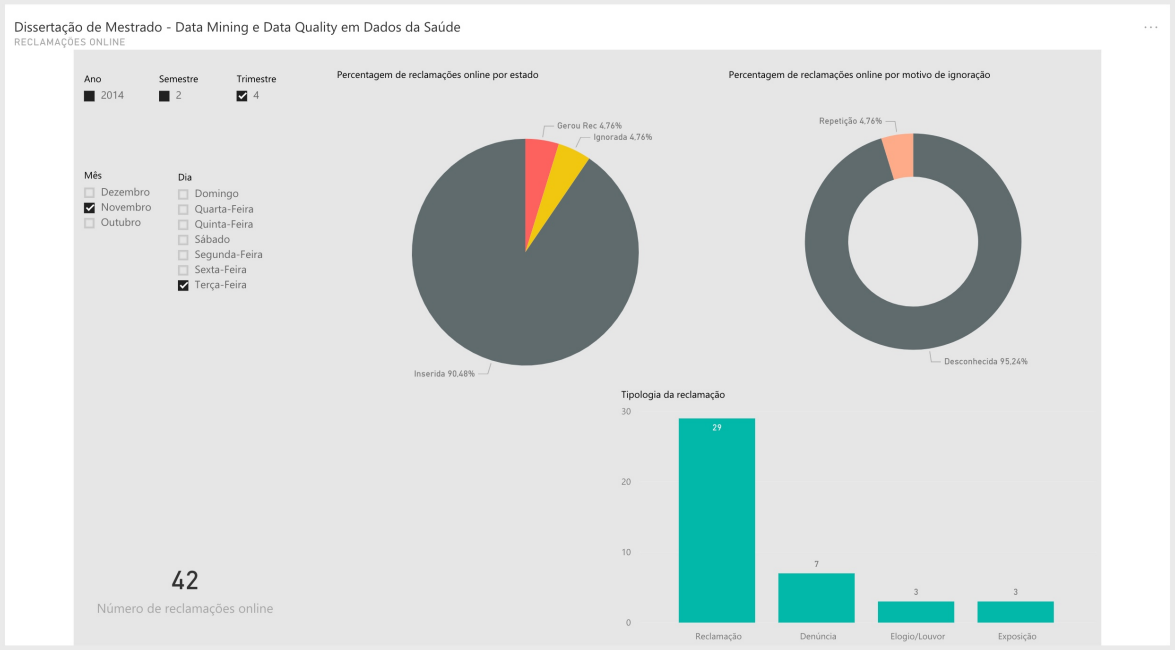

Figura 29 - Análise das reclamações online em 2014 (continuação).

A segunda análise desenvolvida referente ao ano 2015 em que apenas existem registos para o 1º Semestre, conta com um total de 288 reclamações sendo que 126 dizem respeito ao período selecionado, como pode ser visualizado na Figura 30 e na Figura 31. Neste caso foram selecionados dois dias, pois apresentavam igual número de registos. Neste período prevalece:

- Tipificação
    - "Qualidade dos Cuidados de Saúde" com 17,46% dos registos (22 registos);
    - "Taxas moderadoras" com 8,73% dos registos (11 registos);
    - "Direitos dos Utentes" com7,94% dos registos (10 registos);
    - "Tempos de espera" com 7,94% dos registos (10 registos).
- Estado
    - "Inserida" com 99,21% dos registos (125 registos).
- Ignoração
    - "Desconhecida" com 100% dos registos (126 registos)
- Tipologia
    - "Reclamação" com 101 registos.

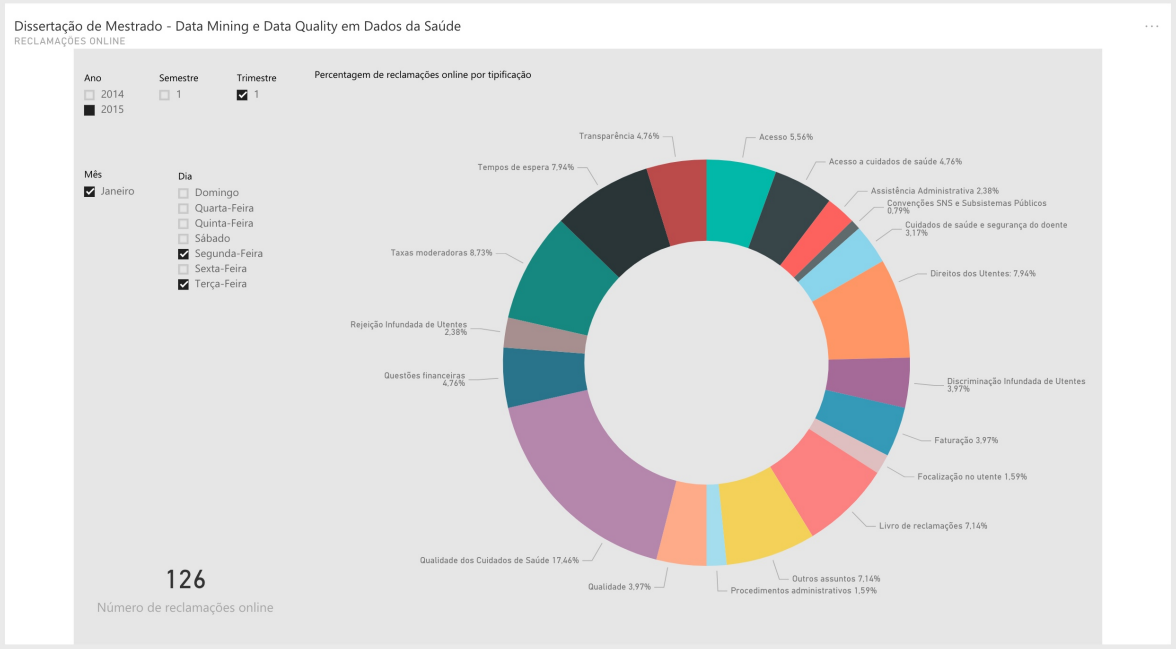

Figura 30 - Análise das reclamações online em 2015.

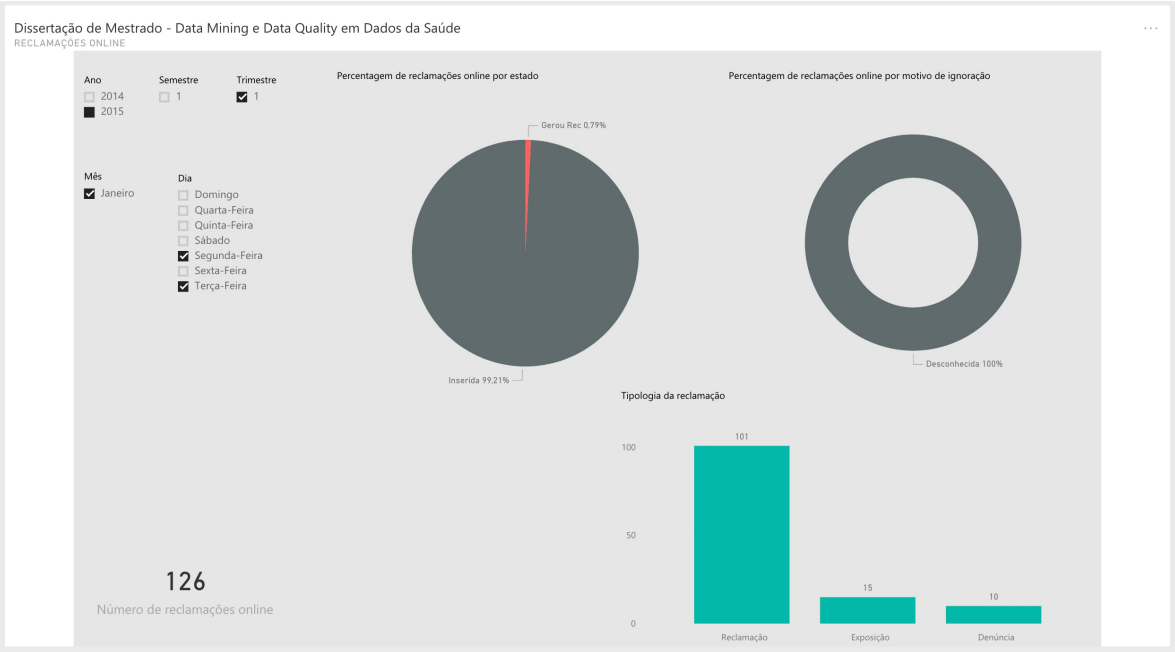

Figura 31 - Análise das reclamações online em 2015 (continuação).

Relativamente às análises das reclamações online apresentadas anteriormente, com base na estação do ano em causa (Inverno) podemos considerar, tal como nas reclamações em papel, que a afluência de reclamações poderá dever-se ao frio e consequência disso a forte ocorrência de gripe. Independentemente da razão da elevada afluência de reclamações nesta altura do ano, podemos observar que a maioria dos registos é do tipo reclamação, que por si só já é motivo de alarme pois algo não agrada a maioria dos utentes. Por outro lado, facilmente identificámos que a maioria das reclamações apresentam o estado inseridas, ou seja, é aceite e analisada. No entanto, caso estas reclamações já inseridas e analisadas sejam ignoradas devem apresentar um motivo. Nesta análise também podemos identificar que a grande maioria dos motivos de ignoração é desconhecida. Perante este caso podemos retirar duas conclusões, podemos considerar que os dados não correspondem e estão incorretos ou podemos considerar falta de profissionalismo no momento de análise. Se estivermos perante o segundo caso e a reclamação seja ignorada sem a identificação de um motivo, o utente não tem forma de entender o que falhou e o que pode mudar para que a reclamação enviada tenha sucesso. Por último, é possível identificar uma clara insatisfação por parte dos utentes no que toca à qualidade dos cuidados de saúde prestados. Subsequentes a esta insatisfação também estão os tempos de espera, as taxas moderadoras e os direitos dos utentes.

É fundamental ter em consideração os factos expostos anteriormente e elaborar atempadamente planos com medidas corretivas, modificando aspetos essenciais para a satisfação do utente. É também importante a elaboração de relatórios para análise da eficácia das medidas tomadas.

5.5 Elaboração de modelos de regressão e classificação

Inicialmente com este projeto de investigação pretendia-se a elaboração de modelos de *Data Mining* capazes de satisfazer necessidades bem definidas anteriormente. Com o desenrolar do projeto, surgiram algumas barreiras que foram originando elevada perda temporal. Após algumas reuniões foram encontradas soluções para o problema encontrado e nesse sentido o projeto atravessou algumas variações em termos de foco. No entanto, após o desenvolvimento da investigação relatada anteriormente foi efetuada uma análise do que poderia ser esperado caso existisse tempo para um desenvolvimento deste tema.

Desta forma, desenvolveram-se duas hipóteses para investigação de *Data Mining* com os mesmos dados, regressão e classificação.

Dando continuidade ao trabalho realizado, relativamente aos modelos de regressão podem ser exploradas as tipificações das reclamações online (Figura 32), com o objetivo principal a previsão mensal do número de ocorrências de cada tipificação. Por outro lado, podem ser exploradas as valências das reclamações em papel (Figura 33), tendo como principal objetivo a previsão mensal do número de ocorrências de cada valência. De modo a responder a estes objetivos é necessária uma preparação prévia dos *datasets*. Tanto para as tipificações como para as valências, a preparação passa pela criação de um ficheiro excel com a contagem por período, neste caso mensal, do número de ocorrências de cada tipificação e de cada valência.

data	tipificacaoid	descricao
04/09/2014	1	Acesso
04/09/2014	1	Acesso
05/09/2014	130	Faturação
05/09/2014	234	Qualidade dos Cuidados de Saúde
06/09/2014	136	Transparência
08/09/2014	267	Outros Assuntos
08/09/2014	1	Acesso
09/09/2014	254	Tempos de Espera
09/09/2014	46	Rejeição Infundada de Utentes
09/09/2014	254	Tempos de Espera
11/09/2014	234	Qualidade dos Cuidados de Saúde
11/09/2014	72	Discriminação Infundada de Utentes
11/09/2014	96	Direitos dos Utentes:
12/09/2014	122	Licenciamento

Figura 32 - Tipificações das reclamações online (excerto).

data	valencia	nome
30/12/2013	28	Medicina Geral e Familiar
02/01/2014	34	Oftalmologia
02/01/2014	28	Medicina Geral e Familiar
02/01/2014	36	Medicina Interna
02/01/2014	11	Cardiologia
02/01/2014	37	Nefrologia
02/01/2014	18	Diálise
02/01/2014	36	Medicina Interna
02/01/2014	11	Cardiologia
02/01/2014	28	Medicina Geral e Familiar
02/01/2014	39	Gastroenterologia
30/12/2013	36	Medicina Interna
02/01/2014	34	Oftalmologia
30/12/2013	35	Atendimento Permanente

Figura 33 - Valências das reclamações em papel (excerto).

Com base na investigação efetuada aos dados podem ocorrer alguns problemas influenciadores do resultado final, como por exemplo:

- Falta de correspondência para todos os dias do mês;
- Existência de muitos valores id_tipificacao nulos.

A estes problemas identificados pode surgir uma questão nem sempre de fácil conclusão, no entanto de bastante importância para dar início à investigação, pois estão sempre pendentes do resultado do problema, como por exemplo:

- Qual a credibilidade do resultado, em termos de aproximação da realidade, se faltam valores?

Relativamente ao desenvolvimento de modelos de classificação foram avaliados vários cenários, sendo que o identificado como mais viável seria a possibilidade de criação de três classes de classificação de ocorrências:

- De 0 a 2 – Razoável;
- De 3 a 6 – Preocupante;
- De 7 a 10 – Crítica.

Perante a distribuição apresentada, seria interessante identificar a área com mais afluência de reclamações, percebendo a gravidade da ocorrência caso esta seja recorrente. O objetivo principal passaria pela rápida identificação de problema, possibilitando atuação imediata na sua resolução.

6 Conclusão

Este documento apresenta o enquadramento conceptual de *Data Science* (DS) e *Data Mining* (DM) no contexto da saúde. Neste enquadramento são apresentados conceitos fundamentais para uma melhor compreensão do tema em estudo como: sistemas de informação, qualidade da informação na saúde, sistema de gestão de qualidade, descoberta de conhecimento em base de dados e sistemas de apoio à decisão. Com base na análise destes conceitos é possível perceber melhor o contexto do problema da qualidade da informação nas reclamações dos serviços prestadores de cuidados de saúde. Desta forma, no desenvolvimento desta revisão literária foi tido em conta o facto de haver necessidade de tratar e analisar os dados fornecidos. Como consequência disso, foi também necessário estudar diversas técnicas de modelação e tratamento de dados. Nesse sentido, foram analisadas técnicas como *Data Mining* e *Data Science* incluindo o *Business Intelligence* e *Big Data*.

Depois de uma investigação aprofundada sobre o tema foi fundamental analisar algumas metodologias que faziam sentido ser utilizadas no desenvolvimento deste projeto, sendo que foram selecionadas três metodologias, duas mais direcionadas à parte de investigação e uma mais direcionada à parte de desenvolvimento: *Case Study, Design Science Research Methodology* e *Kimball Lifecycle*, respetivamente. Do estudo destas metodologias resultou a sua conjugação aplicada em todo o desenvolvimento do projeto.

Ao longo das pesquisas efetuadas, foram encontrados trabalhos relacionados com o uso de *Data Mining* na saúde ou na gestão de reclamações, mas nada em relação à qualidade do processo de análise das reclamações. Deste modo, o facto de ainda ser uma área pouco explorada, despertou ainda mais interesse e curiosidade no desenvolvimento deste projeto.

Posteriormente à revisão literária, numa componente prática foram apresentadas análises primordiais efetuadas aos dados servindo de base para a definição da preparação e tratamento dos mesmos. Neste sentido é apresentado o modelo relacional original e o desenvolvido no contexto deste projeto.

Seguidamente, apresentou-se todo o processo de preparação e tratamento dos dados bem como o processo Extract Transform Load (ETL) e o processamento do cubo *Online Analytical Processing* (OLAP), responsável pela disponibilização dos dados para análise em tempo real.

Com o tratamento dos dados terminado foram apresentadas as análises dos dados desenvolvidas na ferramenta *Power BI*, com o intuito de procurar soluções para a melhoria da qualidade das reclamações na saúde. Através dos *dashboards* apresentados é possível verificar que é necessária uma atuação rápida na estruturação das reclamações tornando-as mais específicas.

Por último, procurando dar resposta ao sucesso de implementação e modelos de *Data Mining* na saúde foram apresentados estudos para desenvolvimentos no que toca a modelos de regressão e classificação.

6.1 Considerações finais

Com o projeto terminado é tempo de fazer uma avaliação referente ao cumprimento dos objetivos definidos inicialmente. Juntamente aos objetivos é importante refletir sobre a questão científica colocada, que procura respostas ao problema apresentado pela Entidade Reguladora da Saúde (ERS).

- *De que modo Data Science pode melhorar a qualidade do processo de análise das reclamações na saúde?*

Neste sentido, como resposta à questão científica, podemos afirmar que através da investigação realizada o *Data Science* pode ajudar a melhorar o processo de análise da qualidade das reclamações, pois com base nos relatórios desenvolvidos é possível observar uma forte incidência de dados sem informação relevante, mostrando a fraca qualidade das reclamações recebidas. Desta forma, é necessário repensar na reestruturação do modelo de reclamação existente, influenciando no processo de especificação dos temas por parte do utente.

Este projeto permitiu atingir o principal objetivo desta dissertação, resultando a apresentação de relatórios automáticos e em tempo real dos dados, tendo por base todo o tratamento e processamento do dataset inicial.

Após o término da investigação e como forma de avaliação do cumprimento dos objetivos, podemos dizer que os resultados obtidos se revelaram satisfatórios.

- A revisão da literatura efetuada permitiu uma boa compreensão de todo o negócio em causa bem como a obtenção de conhecimento acerca de *Data Science* e *Data Mining*;
- A preparação prévia dos dados em estudo bem como o desenvolvimento do processo ETL em ferramentas já conhecidas permitiu uma eficaz análise e tratamento dos dados; Juntamente a este tratamento foi ainda possível obter um processamento eficaz do *dataset*;
- Os dados fornecidos permitiram a compreensão do modelo utilizado, ainda que com alguma dificuldade foi possível uma adaptação do modelo original adequando-o ao tema em estudo;
- Através da criação do cubo OLAP, novamente efetuado em ferramentas já utilizadas em outros projetos, foi possível o desenvolvimento de *dashboards* através do *Power BI*. Ainda que a ligação entre a base de dados e o *Power BI* apresentou alguns problemas, estes foram eficazmente ultrapasados;

- No que toca à criação de modelos de regressão e classificação, devido a problemas encontrados e modificações efetuadas ao projeto, apenas foi possível o desenvolvimento de uma proposta de trabalho, sendo que a sua implementação foi considerada para trabalho futuro.

6.2 Limitações e dificuldades

Ao longo do desenvolvimento deste projeto de dissertação de mestrado foram várias as dificuldades sentidas desde a revisão de literatura até ao desenvolvimento da solução.

Numa primeira fase, na revisão da literatura, foram ultrapassadas dificuldades na procura de documentos científicos relacionados com as reclamações que não incidissem no *text mining*. Outra dificuldade ultrapassada com auxílio dos orientadores remonta para a estruturação da revisão da literatura, pois os objetivos não estavam bem definidos com a entidade parceira de investigação.

Desde início foi difícil o contacto com a ERS, atrasando o projeto em todas as suas etapas. A estes problemas de comunicação alinhava a incerteza da definição de objetivos e requisitos comprometendo a viabilidade do projeto. Inúmeras foram as tentativas de contacto pedidas aos orientadores, mas sempre não correspondidas pela parte da ERS. Estas dificuldades foram sendo ultrapassadas com base na experiência dos orientadores na área, definindo desta forma os requisitos e objetivos.

Numa segunda fase, já no desenvolvimento da solução, ainda sem qualquer resposta da ERS deparamo-nos com o não fornecimento dos dados para investigação. Esta foi a dificuldade mais marcante no projeto, porque sem os dados para investigação comprometia o desenvolvimento de todo o projeto. No sentido de resolução urgente deste problema, em conjunto com os orientadores após várias tentativas de contacto foi estipulado um prazo de resposta, em que se não houvesse feedback a solução a desenvolver teria de ser repensada. Esta dificuldade foi ultrapassada com a disponibilização de dados por parte dos orientadores, no entanto todo o desenvolvimento teve de sofrer alterações bem como os objetivos e solução a desenvolver. No início do projeto de investigação estava previsto a criação de modelos de *Data Mining* seguindo a metodologia *Crisp-DM,* mas em virtude dos atrasos verificados e falta de resposta por parte da ERS foi impossível o desenvolvimento dos mesmos, sendo apenas apresentado na secção 5.5 uma possível proposta estudada para desenvolvimento futuro.

Numa perspetiva mais técnica, foi sentida alguma dificuldade e incompatibilidade no manuseamento de ferramentas *microsoft* no *MacOS*, sendo facilmente ultrapassada através de algumas pesquisas.

Além de todas estas dificuldades mencionadas anteriormente é importante referir que foi difícil conciliar a complexidade do projeto de investigação com a carga e complexidade dos projetos da empresa laboral, tendo sido a carga horária um fator problemático.

6.3 Análise de riscos

Num projeto deste tipo é essencial uma identificação atempada de riscos bem como de estratégias de atenuação dos efeitos dos mesmos. Deste modo, a Tabela 10 considerada em toda a execução deste projeto representa uma identificação de todos os riscos que diretamente afetaram o desenvolvimento do projeto. Na Tabela 10 é ainda possível visualizar a descrição dos riscos identificados, a probabilidade de acontecimento, o impacto, a ação atenuante e a verificação da ocorrência do risco.

Tabela 10 - Lista de Riscos

Identificação do Risco	Probabilidade	Impacto	Seriedade[4]	Descrição do Impacto	Ação Atenuante	Risco Verificado
Atraso na disponibilização dos dados	4	5	20	Muito alto	Ocorrência de reuniões com os orientadores onde a solução encontrada foi a disponibilização de dados que os mesmos possuíam de estudos passados.	Sim
Complexidade do Projeto	4	4	16	Alto	Divisão do projeto em etapas e tarefas para o desenvolvimento do projeto com base nos requisitos alterados. Desenvolvimento efetuado sobre o auxílio de três metodologias.	Sim
Má qualidade dos dados	4	4	16	Muito alto	Consulta dos orientadores e de trabalhos relacionados para auxílio na compreensão dos dados. Pesquisas adicionais relativas a termos técnicos utilizados na medicina.	Sim

[4] Seriedade = Probabilidade x Impacto

6.4 Trabalho futuro

Após conclusão deste projeto de investigação, existem alguns pontos a serem desenvolvidos no futuro, destacando-se entre eles:

- Adição de atributos de estudo ao modelo já desenvolvido, alargando a exploração de indicadores através de desenvolvimento de *dashboards*.
- Desenvolvimento e aplicação dos modelos de regressão e classificação mencionadas na secção 5.5;
- Desenvolvimento do mesmo estudo, mas com base nos dados atualizados da ERS;
- Elaboração de artigo científico referente ao estudo efetuado neste projeto de investigação.

7 Referências Bibliográficas

Almeida, L. (2010). A criação da Entidade Reguladora da Saúde em Portugal | Comunicação Por conta e Risco. Retrieved October 18, 2016, from https://porcontaerisco.wordpress.com/2010/12/03/a-criacao-da-entidade-reguladora-da-saude-em-portugal/#comments

Almeida, M., & Bax, M. (2003). Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção.

Alves, D. da S. (2015). *Saúde em Portugal: Estudo das Urgências Hospitalares através do Data Mining*.

Anabela, L. (2014). *Entre o direito a reclamar e o direito à saúde. Serviço social em gabinetes do cidadão, do SNS*.

Berner, E. S. (1999). *Clinical decision support systems: theory and practice*. *Biomedical Informatics*. Retrieved from http://books.google.com/books?hl=en&lr=&id=2iqBv92loHoC&oi=fnd&pg=PR7&dq=Clinical+Decision+Support+systems+theory+and+Practice&ots=mGvFZt_Gm7&sig=McIidMt6NDu9eY5vPwOQFA-nMVU

Berry, M. J. A., & Linoff, G. S. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Second Edition*.

Berson, A., & Smith, S. J. (1997). *Data warehousing, data mining, and OLAP*. McGraw-Hill.

Carvalho, J. A. (2000). Information System? Which One Do You Mean?

Chen, M., Mao, S., Liu, Y., Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. http://doi.org/10.1007/s11036-013-0489-0

Ciclo PDCA. (n.d.). Retrieved February 18, 2017, from http://npu.com.br/wp-content/uploads/2016/03/CICLO.png

Entidade Reguladora da Saúde. (2016). Entidade Reguladora da Saúde. Retrieved October 18, 2016, from https://www.ers.pt/pages/2

Espanha, R. (2010). Adenda à Análise Especializada: Tecnologias de Informação e Comunicação.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34. http://doi.org/10.1145/240455.240464

Ferreira, A. S. (2004). Do que falamos quando falamos de regulação em saúde?**. *Análise Social*, (171), 313–337. Retrieved from

http://analisesocial.ics.ul.pt/documentos/1218705541T1vTG4xh0Gk67XU7.pdf

Ferreira, M., Reis, L. P., Gonçalves, J., & Rocha, Á. (2015). Data Mining e Sistemas de Apoio à Decisão em Aplicações Clínicas e Qualidade de Vida Data Mining and Decision Support Systems for Clinical Application and Quality of Life.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases: An Overview.

Freitas, M. G. (2015). Comunicado do Diretor-Geral da Saúde: Época de gripe 2014/2015. *Direção Geral Da Saúde*, 1–2. Retrieved from www.dgs.pt/a-direccao-geral.../epoca-de-gripe-20142015-pdf2.aspx

Freixo, J., & Rocha, Á. (2014). Arquitetura de Informação de Suporte à Gestão da Qualidade em Unidades Hospitalares Information Architecture to Support Quality Management in Hospital Units. http://doi.org/10.17013/risti.14.1-15

Gagliardi, A., & Jadad, A. R. (2002). Examination of instruments used to rate quality of health information on the internet: chronicle of a voyage with an unclear destination. *BMJ*, *324*(7337).

Goebel, M., & Gruenwald, L. (1999). A SURVEY OF DATA MINING AND KNOWLEDGE DISCOVERY SOFTWARE TOOLS.

Han, J., & Kamber, M. (1998). *Data Mining Concepts and Techniques*.

Johansson, R. (2003). Case Study Methodology.

Krishnan, K. (2013). *Data Warehousing in the Age of Big Data*.

Marreiros, M. (2007). *Agentes de Apoio à Argumentação e Decisão em Grupo*. Retrieved from https://repositorium.sdum.uminho.pt/bitstream/1822/7643/1/Tese_Maria Goreti Carvalho.pdf

Microsoft. (2010). What is Power BI | Microsoft Power BI. Retrieved October 17, 2017, from https://powerbi.microsoft.com/en-us/what-is-power-bi/

Moreira, A. (2002). Uso de ontologia em sistemas de informação computacionais. *Perspectivas Em Ciência Da Informação*, *7*(1), 49–60. Retrieved from http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/413

Moreira, V. (2011). Regulação dos serviços da saúde - público. Retrieved from https://www.publico.pt/opiniao/jornal/regulacao-dos-servicos-da-saude-23006123

Navega, S. (2002). Princípios Essenciais do Data Mining.

Negash, S. (2004). Business intelligence. *The Communications of the Association for Information*, *13*(15), 177–195. http://doi.org/10.1007/978-3-540-48716-6_9

Nemati, H. R., & D. Barko, C. (2010). Organizational Data Mining. In *Data Mining and Knowledge*

Discovery Handbook (pp. 1041–1048).

Oded, M., & Rokach, L. (2010). Introduction to Knowledge Discovery and Data Mining. In *Data Mining and Knowledge Discovery Handbook* (pp. 1–15). http://doi.org/10.1007/0-387-25465-x_2

Oliveira, A. A. G. da S. (2015). *Apoio à Decisão na Análise Inteligente de Reclamações*. Universidade do Minho.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. http://doi.org/10.2753/MIS0742-1222240302

Pereira, J. (2005). *Modelos de Data Mining para multi-previsão: aplicação à medicina intensiva*.

Pomerol, J.-C., & Adam, F. (2004). Practical Decision Making – From the Legacy of Herbert Simon to Decision Support Systems.

Provost, F., & Fawcett, T. (2013). DATA SCIENCE AND ITS RELATIONSHIP TO BIG DATA AND DATA-DRIVEN DECISION MAKING. Retrieved from http://online.liebertpub.com/doi/pdf/10.1089/big.2013.1508

Ramos, I., & Santos, M. Y. (2003). Data Mining no suporte à construção de Conhecimento Organizacional.

Reis, C. (2011). *Modelos de Gestão Hospitalar*. UNIVERSIDADE DE COIMBRA FACULDADE DE ECONOMIA.

Ribeiro, V. (2011). O que é Data Warehouse? Retrieved from https://vivianeribeiro1.wordpress.com/2011/03/30/o-que-e-data-warehouse/

Ross, M. (2009). Design Kimball Lifecycle - Kimball Group. Retrieved January 29, 2017, from http://www.kimballgroup.com/2009/08/design-tip-115-kimball-lifecycle-in-a-nutshell/

Sandi, A. A. A. (2015). *A importância dos Sistemas de Informação em Saúde – Estudo de caso na USF CelaSaúde*.

SGS. (2017). SGS Portugal - ISO 9001 - Certificação - Sistemas de Gestão da Qualidade - Saúde & Segurança. Retrieved January 22, 2017, from http://www.sgs.pt/pt-PT/Health-Safety/Quality-Health-Safety-and-Environment/Quality/Quality-Management-Systems/ISO-9001-Certification-Quality-Management-Systems.aspx

Simões, J. (2004). As parcerias público-privadas no sector da saúde em Portugal, *4*.

Simon, H. (1977). *The New Science of Management Decision*.

SPMS. (2017a). PEM - SPMS. Retrieved January 22, 2017, from http://spms.min-saude.pt/product/pem/

SPMS. (2017b). SClínico - Cuidados de Saúde Primários (CSP) - SPMS. Retrieved January 22, 2017,

from http://spms.min-saude.pt/product/sclinicocsp/

Turban, E., E. Aronson, J., & Liang, T.-P. (2007). Decision Support Systems and Business Intelligence. In *Decision Support and Business Intelligence Systems, 7/E* (pp. 1–35). http://doi.org/10.1017/CBO9781107415324.004

Vasconcelos Parra, R. (2014). *Reclamações no setor público da saúde.*

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making* (Vol. 53). http://doi.org/10.1017/CBO9781107415324.004

Zainal, Z. (2007). Case study as a research method. *Jurnal Kemanusiaan Bil*, *9*.

ANEXO I – DIAGRAMA DE *GANTT*

Nome da Tarefa	Duração	Início	Conclusão
Data Mining e Data Quality em Dados da Saúde	**291 dias**	**Ter 20-09-16**	**Ter 31-10-17**
Reunião de Orientação	0 dias	Ter 20-09-16	Ter 20-09-16
Reunião de Orientação	0 dias	Sex 21-10-16	Sex 21-10-16
Reunião de Orientação	0 dias	Qua 26-10-16	Qua 26-10-16
Reunião de Orientação	0 dias	Sex 10-02-17	Sex 10-02-17
Reunião de Orientação	0 dias	Qui 02-03-17	Qui 02-03-17
Reunião de Orientação	0 dias	Sex 31-03-17	Sex 31-03-17
Reunião de Orientação	0 dias	Sex 28-04-17	Sex 28-04-17
Reunião de Orientação	0 dias	Sex 02-06-17	Sex 02-06-17
Reunião de Orientação	0 dias	Seg 31-07-17	Seg 31-07-17
Reunião de Orientação	0 dias	Sex 01-09-17	Sex 01-09-17
Reunião de Orientação	0 dias	Seg 23-10-17	Seg 23-10-17
Etapa 1	**115 dias**	**Ter 20-09-16**	**Seg 27-02-17**
Identificação do problema e motivaçã	4 dias	Qua 21-09-16	Sáb 24-09-16
Definição dos requisitos de negócio	22 dias	Sáb 24-09-16	Seg 24-10-16
Definição de objetivos	22 dias	Sáb 24-09-16	Seg 24-10-16
Planeamento do Projeto	89 dias	Qua 26-10-16	Seg 27-02-17
Compreensão do negócio	89 dias	Qua 26-10-16	Dom 26-02-17
Argumentação lógica	113 dias	Qua 21-09-16	Sex 24-02-17
Interpretação histórica	113 dias	Qua 21-09-16	Sex 24-02-17
Entrega da Pré-dissertação	1 dia	Seg 27-02-17	Seg 27-02-17
Etapa 2	**115 dias**	**Ter 28-02-17**	**Dom 06-08-17**
Compreensão dos dados	21 dias	Ter 28-02-17	Ter 28-03-17
Preparação dos dados	49 dias	Qua 01-03-17	Sáb 06-05-17
Design e Conceção	10 dias	Dom 07-05-17	Qui 18-05-17
Desing e Seleção da Arquitetura	58 dias	Qui 18-05-17	Dom 06-08-17
Design e Desenvolvimento de aplicações BI	58 dias	Qui 18-05-17	Dom 06-08-17
Qualitativo	115 dias	Ter 28-02-17	Dom 06-08-17
Correlacional	115 dias	Ter 28-02-17	Dom 06-08-17
Etapa 3	**126 dias**	**Sáb 25-03-17**	**Sex 15-09-17**
Modelação	97 dias	Sáb 25-03-17	Dom 06-08-17
Design e Conceção	10 dias	Dom 07-05-17	Qui 18-05-17
Design e Seleção da Arquitetura	58 dias	Qui 18-05-17	Dom 06-08-17
Desenvolvimento do ETL e Modelação Dimensional	58 dias	Qui 18-05-17	Dom 06-08-17
Design e Desenvolvimento de aplicações BI	58 dias	Qui 18-05-17	Dom 06-08-17
Experimental	31 dias	Dom 06-08-17	Sex 15-09-17
Etapa 4	**30 dias**	**Sex 15-09-17**	**Qui 26-10-17**
Demonstração	17 dias	Sex 15-09-17	Dom 08-10-17
Simulação	11 dias	Dom 08-10-17	Sex 20-10-17
Avaliação	5 dias	Sex 20-10-17	Qui 26-10-17
Comunicação	151 dias	Dom 26-03-17	Sex 20-10-17
Etapa 5	**23 dias**	**Dom 01-10-17**	**Ter 31-10-17**
Implementação e Manutenção	16 dias	Dom 01-10-17	Sex 20-10-17
Revisão da Dissertação	7 dias	Sex 20-10-17	Seg 30-10-17
Entrega Final	1 dia	Ter 31-10-17	Ter 31-10-17